University of Alberta

APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS IN DRINKING WATER TREATMENT PROCESS MODELLING AND CONTROL

By

Christopher Wayne Baxter

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**

in

Environmental Science

Department of Civil and Environmental Engineering

Edmonton, Alberta

Spring 2002

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-68542-X

Canada

University of Alberta

Library Release Form

Name of Author: Christopher W. Baxter

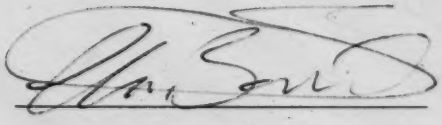
Title of Thesis: Applications of Artificial Neural Networks in Drinking Water Treatment Process Modelling and Control

Degree: Doctor of Philosophy

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



Christopher W. Baxter
155 Brookhaven Ave.
Dorval, Quebec
H9S 2N5

Date: January 31, 2002

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis titled **Applications of Artificial Neural Networks in Drinking Water Treatment Process Modelling and Control** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Environmental Science**.

D.W. Smith, Ph.D., P.Eng.
Co-supervisor

S.J. Stanley, Ph.D., P.Eng.
Co-supervisor

W.B. Kindzierski, Ph.D., P.Eng.
Committee Chair

I.D. Buchanan, Ph.D., P.Eng.
Committee Member

J.J. Leonard. Ph.D., P.Eng.
Committee Member

B.J. Lence, Ph.D., P.Eng.
External Examiner

January 31, 2002
Date

This work is dedicated to my amazing wife and best friend Sheri Patterson, whose continual love, support, and friendship has enabled me to truly enjoy life, and to my family for supporting and taking pride in my academic achievements.

ABSTRACT

The future of the drinking water treatment industry will be characterized by a more stringent regulatory environment as well as increased pressures to provide efficient treatment. As such, utilities must actively search out new technologies that allow for a greater understanding of process operations and improved process optimization. When applied to drinking water treatment process modelling and control, artificial neural networks (ANNs) have the potential to significantly reduce operational and maintenance costs, improve customer service, and improve water quality.

The primary objective of the current study is to develop ANN models for various treatment process parameters at pilot-scale and full-scale treatment facilities, owned and operated by the Metropolitan Water District of Southern California and EPCOR Water Services, and subsequently apply the models in process assessment, optimization, and control. In the process of meeting this objective, many important aspects related to ANN-based process modelling and control at drinking water treatment facilities are investigated. More specifically, the quantity and quality of data required for successful model development, the quantification of model prediction boundaries, and the limitations of the ANN technology are addressed

This research contributes significant new knowledge to the field of drinking water treatment operations. With respect to process modelling, the research will result in the creation of a robust protocol for developing ANN models of water treatment processes. The development of ANN-based offline and online process control and optimization

applications will provide plant operators with new and powerful options for scenario analysis, training, decision verification, and virtual experimentation. The application of ANNs to the analysis of pilot-scale data will provide experimenters with a simplified methodology for gaining important insight into the impacts of key factors on process operations. Finally, the creation of an operational pilot-scale model-based advanced process control system will allow for automated optimization of chemical dosing and other operational characteristics.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial and organizational support of EPCOR Water Services, the National Sciences and Engineering Research Council of Canada, the American Water Works Association Research Foundation, and the Killam Trusts. In addition, several individuals deserve special recognition for their contributions:

- Daniel W. Smith and Stephen J. Stanley, my supervisors, for providing a truly enriching and well-rounded academic experience, and for encouraging and supporting my participation in conferences, teaching, consulting, and extra-curricular activities
- Riyaz Shariff, Qing Zhang, and Evan Saummer, for helping to advance the ANN technology through the implementation of new ideas
- Charley Hartery of EPCOR Water Services for providing operational guidance at the EPCOR pilot plant and for his assistance in data collection
- Kevin Graff from the Metropolitan Water District of Southern California, for his assistance in data collection and troubleshooting
- Craig Bonneville, Simon Thomas, and Christian Madsen of EPCOR Water Services for answering countless process operations questions and providing invaluable feedback on model applications

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1. Background 1
1.1.1. General Characteristics of Artificial Neural Networks 2
1.1.2. Existing ANN Models in the Water Treatment Industry 4
1.1.3. Integration of ANN Models for Process Optimization and Control 5
1.2. Research Needs 6
1.3. Objectives 7
1.3.1. Creation of a Model Development Protocol 8
1.3.2. Modelling at MWD Facilities 8
1.3.3. Virtual Laboratory Experimentation and Scenario Analysis 9
1.3.4. Defining Model Boundaries 10
1.3.5. Analysis of Pilot-scale Data 10
1.3.6. Advanced Process Control 11
1.4. Organization 11
1.5. References 12
2. DEVELOPING ARTIFICIAL NEURAL NETWORK MODELS OF DRINKING WATER TREATMENT PROCESSES: A GUIDE FOR UTILITIES 15
2.1. Introduction 15
2.2. The Artificial Neural Network Technology 16
2.2.1. Types of ANN Models 17
2.2.2. Key Components of ANN Models 18
2.2.3. ANN Learning 19
2.2.4. Advantages of ANN Modelling 21
2.2.5. Challenges of ANN Modelling 22
2.2.6. ANN Applications 22
2.2.7. Existing ANN Models 23
2.3. Building ANN Models of Drinking Water Treatment Processes 24
2.3.1. Needs and Suitability Assessment 24
2.3.2. Data Collection and Analysis 26
2.3.3. Application of the Model Building Protocol 28

2.3.3.1. Selection of Model Inputs and Outputs 29
2.3.3.2. Selection and Organization of Data Patterns 30
2.3.3.3. Determination of Architecture Characteristics 31
2.3.3.4. Evaluation of Model Stability (Cross-Validation) 32
2.3.3.5. Model Fine-tuning 33
2.3.3.6. Repetition of Modelling Protocol Steps 33
2.3.4. Performance Evaluation 34
2.4. Advanced Modelling 36
2.5. Conclusion 37
2.6. References 38
3. PROCESS MODELLING AND MODEL APPLICATIONS FOR METROPOLITAN WATER DISTRICT OF SOUTHERN CALIFORNIA FACILITIES 43
3.1. Introduction 43
3.2. Background Information 43
3.2.1. Oxidation Demonstration Project Plant 43
3.2.2. F.E. Weymouth Filtration Plant 44
3.3. ANN Model Development 45
3.3.1. Methodology 45
3.3.1.1. Data Collection and Management 45
3.3.1.2. Software 46
3.3.1.3. Model Development Protocol 47
3.4. Results 48
3.4.1. Source Data Analysis 48
3.4.1.1. Oxidation Demonstration Project Plant. 48
3.4.1.2. F.E. Weymouth Filtration Plant. 50
3.4.2. Model Development and Evaluation 51
3.4.2.1. Oxidation Demonstration Project Plant. 51
3.4.2.2. F.E. Weymouth Filtration Plant. 56
3.5. Model Applications 60
3.5.1. Virtual Laboratory and Scenario Analysis 61
3.5.1.1. Methodology and Results 61

3.5.1.1.1. Virtual Laboratory Applications 61
3.5.1.1.2. Scenario Analysis Applications 64
3.5.1.1.3. Operator Training 65
3.6. Conclusions 66
3.7. References 67
4. EVALUATION OF MODEL BOUNDARIES 81
4.1. Introduction 81
4.2. Background Information 82
4.2.1. Expanded Data Domain Evaluation 82
4.2.2. Evaluation of Scaling Effects 83
4.3. Methodology 84
4.3.1. Expanded Data Domain Evaluation 84
4.3.2. Evaluation of Scaling Effects 86
4.4. Results and Discussion 87
4.4.1. Expanded Data Domain Evaluation 87
4.4.1.1. Out-of-range Inputs 87
4.4.1.2. New Water Quality and Operational Data 92
4.4.2. Evaluation of Scaling Effects 94
4.5. Conclusion 97
4.6. References 98
5. USING ARTIFICIAL NEURAL NETWORKS TO ANALYZE PILOT-SCALE DATA 107
5.1. Introduction 107
5.2. The ANN Technology 109
5.3. Multiple Regression Analysis 110
5.3.1. General Characteristics of Multiple Regression Analyses 110
5.3.2. Assumptions Implicit in Multiple Regression 111
5.3.3. Checking the Adequacy of the Model 112
5.4. Methods 114
5.5. Results and Applications 115
5.5.1. ODP Plant Analysis 115

5.5.1.1. Multiple Regression Model Results 116
5.5.1.2. ANN Model Results 119
5.5.1.3. ANN Model Applications 120
5.5.2. EPCOR Water Services Pilot Plant Analysis 121
5.5.2.1. Multiple Regression Model Results 122
5.5.2.2. ANN Model Results 124
5.5.2.3. ANN Model Applications 125
5.6. Limitations of the Technique 126
5.7. Conclusions 127
5.8. References 128
6. MODEL-BASED CONTROL OF ENHANCED COAGULATION 140
6.1. Introduction 140
6.2. Background Information 141
6.2.1. EPCOR Water Services Pilot Plant 141
6.2.2. The Coagulation Process 142
6.2.3. ANN Modelling 143
6.2.4. Model-based Advanced Process Control 144
6.2.4.1. Control Logic 144
6.2.4.2. Online Analyzers 145
6.2.4.3. SCADA System 146
6.2.5. Model Integration 146
6.3. Methods 147
6.3.1. Data Collection 147
6.3.2. ANN Model Development and Control System Integration 148
6.4. Results and Discussion 148
6.4.1. ANN Model Development and Evaluation 148
6.4.2. Control System Integration 150
6.4.3. Control System Evaluation 152
6.5. Conclusions 157
6.6. References 158
7. GENERAL DISCUSSION AND CONCLUSISONS 164

7.1. Introduction 164
7.2. Discussion and Application Potential 165
7.2.1. Process Assessment and Data Analysis 165
7.2.2. Offline Operational Tools 168
7.2.3. Online Operational Tools 169
7.2.4. Advanced Process Control 170
7.3. Summary of Major Findings 171
7.4. Recommendations for Future Study 172
7.5. Conclusion 174
APPENDIX A – MODELLING DATA 176
APPENDIX B – CURRICULUM VITAE 182

LIST OF TABLES

Table 2.1 Common factors and values in determining ANN architecture characteristics 41
Table 3.1 Water quality variables measured during the filtration study at the ODP Plant 68
Table 3.2 Operational variables measured during the filtration study at the ODP Plant .. 68
Table 3.3 Water quality variables measured at the F.E. Weymouth Filtration Plant 69
Table 3.4 Operational variables measured at the F.E. Weymouth Filtration Plant 69
Table 3.5 Variables measured by online instruments at the F.E. Weymouth Filtration Plant 70
Table 3.6 ODP Plant, data analysis of raw water quality variables 70
Table 3.8 F.E. Weymouth Filtration Plant, data analysis of raw water quality variables 71
Table 3.9 F.E Weymouth Filtration Plant, data analysis of operational and filter effluent variables 72
Table 3.10 ODP Plant, model input variables 72
Table 3.11 ODP Plant, modelling results 72
Table 3.12 ODP Plant, model cross-validation results 73
Table 3.13 F.E. Weymouth Filtration Plant, model input variables 73
Table 3.14 F.E. Weymouth Filtration Plant, modelling results 73
Table 3.15 F.E. Weymouth Filtration Plant, model cross-validation results 73
Table 3.17 ODP Plant, scenario analysis event data 74
Table 4.1 Boundaries between modelling and expanded domain data sets 99
Table 4.2 Expanded data domain model results 99
Table 4.3 95% confidence intervals for the expanded domain colour model regression equations 99
Table 4.4 Kohonen ANN classification results for clarifier effluent colour model data 100
Table 4.5 Model results for winter and summer category models 100
Table 4.6 Category 1 (winter) model results when applied to other categories 100

Table 4.7 Category 4,6, and 7 (summer) model results when applied to other categories 100
Table 5.1 Levels of fixed factors investigated during the ODP study 130
Table 5.2 Statistical summary of raw water quality variables during the ODP Plant study 130
Table 5.3 ODP Plant model variables 130
Table 5.4 ODP Plant multiple regression model coefficients 131
Table 5.5 Statistical summary of data collected during the EPCOR Pilot Plant study ... 131
Table 6.1 ANN model variables for the advanced process control system 159
Table 6.2 ANN model results for the advanced process control system 159
Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3 177

LIST OF FIGURES

Figure 2.1 The main stages of developing an ANN process model 42
Figure 3.1 ODP Plant, raw water quality variables 75
Figure 3.2 F.E. Weymouth Filtration Plant, 50th percentile plant influent particle counts 75
Figure 3.3 ODP Plant, 50th percentile filter effluent turbidity model results 76
Figure 3.4 ODP Plant, 50th percentile filter effluent particle counts model results 76
Figure 3.5 F.E. Weymouth Filtration Plant, combined filter effluent turbidity model results 77
Figure 3.6 F.E. Weymouth Filtration Plant, mean filter effluent particle counts model results 77
Figure 3.7 F.E. Weymouth Filtration Plant, effect of raw water turbidity and particle counts on filter effluent particle counts for typical summer water 78
Figure 3.9 F.E. Weymouth Filtration Plant, effect of alum and polymer doses on filter effluent particle counts for typical summer water 79
Figure 3.10 F.E. Weymouth Filtration Plant, effect of plant flow and filtration rate on filter effluent particle counts for typical summer water 79
Figure 3.11 ODP Plant August 26th 1997, scenario analysis for chemical dose optimization 80
Figure 3.12 F.E. Weymouth Filtration Plant September 10th 1999, optimization of filter performance through plant flow variation 80
Figure 4.1 Rossdale WTP, expanded domain model predictions on production data set 101
Figure 4.2 Determination of the relationship between raw water turbidity and absolute prediction error 101
Figure 4.3 E.L. Smith WTP, model results for the clarifier effluent colour model (75th percentile value for raw water colour boundary) 102
Figure 4.4 Impact of raw water colour on prediction error (75th percentile boundary) .. 102
Figure 4.5 Determination of the relationship between raw water colour and absolute prediction error (75th percentile boundary) 103

Figure 4.6 Impact of raw water colour on prediction error (95th percentile boundary).. 103
Figure 4.7 Determination of the relationship between raw water colour and absolute prediction error (95th percentile boundary) 104
Figure 4.8 Category 1 model results when applied to Category 2 data 104
Figure 4.9 Category 1 model results when applied to Category 3 and Category 4 data. 105
Figure 4.10 Effect of scaling variables on expanded domain predictions 105
Figure 4.11 Generation of erroneous predictions using a model with open scaling..... 106
Figure 5.1 ODP Plant filter effluent particle counts regression model results 132
Figure 5.2 ODP Plant filter effluent turbidity regression model results 132
Figure 5.3 ODP Plant filter effluent turbidity regression model, normal plot of residuals 133
Figure 5.4 ODP Plant filter effluent particle counts regression model, model residuals against observed values..... 133
Figure 5.5 ODP Plant filter effluent particle counts ANN model results 134
Figure 5.6 ODP Plant filter effluent turbidity ANN model results..... 134
Figure 5.7 The effects of alum dose and polymer dose on filter effluent turbidity at the ODP Plant 135
Figure 5.8 The effect of plant flow and influent temperature on filter effluent particle counts at the ODP Plant 136
Figure 5.9 EPCOR pilot plant clarifier effluent turbidity multiple regression model results 137
Figure 5.10 EPCOR Pilot Plant clarifier effluent turbidity ANN model results 137
Figure 5.11 EPCOR pilot plant ANN model, effect of influent turbidity and alum dose on clarifier effluent turbidity during spring break-up conditions 138
Figure 5.12 EPCOR Pilot plant ANN model, effect of influent turbidity and influent colour on clarifier effluent turbidity during typical summer operations..... 139
Figure 6.1 Information flow between advanced process control system components ... 160
Figure 6.2 Schematic diagram of the EPCOR Pilot Plant coagulation process..... 160
Figure 6.3 Simplified version of the control system algorithm 161
Figure 6.4 Case 1 advanced process control system results 162
Figure 6.5 Case 3 advanced process control system results 162

Figure 6.6 Case 5 advanced process control system results 163

LIST OF SYMBOLS AND ABBREVIATIONS

β_i	true slope of the regression surface in the x_i direction
β_o	true Y-intercept
°C	degrees centigrade
$	Canadian dollars
ε	regression error term.
#	number
# Obs.	number of observations
%	percent
% SPW	percent State Project Water
% v/v	percent volume/volume
AI	artificial intelligence
ANN	artificial neural networks
ANOVA	analysis of variance
b_i	true slope estimate
$CaCO_3$	calcium carbonate
cm	centimeter(s)
counts/mL	particle counts per milliliter
confid.	confidence
COV	coefficient of variation
CRW	Colorado River water
deg.	degrees
DLL	dynamic link library
ft	feet
gpm/ft^2	gallons per minute/square feet
GUI	graphical user interface
h	hour(s)
IMC	internal model control
inf.	influent
IPS	intelligent problem solver
Ln	natural logarithm
M	megabytes
m^3	cubic meters
MAE	mean absolute error

MAPE	mean absolute percent error
m^3/h	cubic meters/hour
MGD	million gallons per day
mg/L	milligrams per liter
MHz	megahertz
MIMO	multiple input multiple output
Min	minimum
Max	maximum
MWD	Metropolitan Water District of Southern California
NTU	nephelometric turbidity units
ODP	Oxidation Demonstration Project
pct.	percentile
PLC	programmable logic controller
r	coefficient of correlation
R^2	coefficient of multiple determination
RAM	random access memory
SCADA	supervisory control and data acquisition
SI	International System of Units
SNN	Statistica Neural Networks
SPW	State Project Water
SSE	sum of squares, error
SST	sum of squares, total
Std. Dev.	standard deviation
std. error	standard error
TCU	true colour units
THM	trihalomethane(s)
TON	total odour number
µm	micrometers
UV_{254}	ultraviolet-254 absorbance
WTP	water treatment plant
x_i	independent variable
Y	dependent variable
y_i	observed value of the dependent variable
$\hat{y}_i$	predicted value of the dependent variable

1. INTRODUCTION

1.1. BACKGROUND

The regulatory environment in the drinking water treatment industry is becoming increasingly complex due to constantly evolving treatment technologies and more stringent standards for the removal of chemical, physical, and biological contaminants. Unit processes in the water treatment industry are characterized by complex non-linear relationships between numerous process input and output variables. Historically, attempts have been made to model these relationships by fitting data to severely constrained empirical models based on bench-scale or pilot-scale data. Such attempts have generally been unable to account for simultaneous variations in process variables and have therefore exhibited sub-optimal performance when applied to full-scale systems. As utilities strive to respond to both regulatory pressures, as well as those imposed by an increasingly savvy and demanding customer base, new technologies for process optimization and control are needed.

With the recent developments in computing systems, artificial intelligence (AI) techniques have been used and show promise for treatment modelling, control, and optimization. One AI technology that is particularly suited for water treatment problems is the artificial neural network (ANN) technology, which focuses on finding repeated, recognizable, and predictable patterns between model input and output variables. The current study focuses on the application of this technology to drinking water treatment

process modelling and control. As such, the remainder of this section elaborates on the applicability of the technology to the water treatment industry.

1.1.1. General Characteristics of Artificial Neural Networks

The ANN technology is one that uses artificial intelligence in an attempt to mimic the human brain's problem solving capabilities. ANNs are capable of self-organization and learning; patterns and concepts can be extracted directly from historical data (Baxter, Stanley, and Zhang 1999). In general, artificial neural networks can be applied to the following types of problems: pattern classification, clustering and categorization, function approximation, prediction and forecasting, optimization, associative memory, and process control (Jain, Mao, and Mohuuddin 1996). When presented with data patterns, sets of historical input and output data that describe the problem to be modelled, ANNs map the cause-effect relationships between the model inputs and outputs. This mapping of input/output relationships in the ANN model architecture allows developed models to be used to predict the value of the model output variable, given any reasonable combination of model input data, with satisfactory accuracy.

The ANN modelling technique holds several advantages over mechanistic modelling that make it particularly suitable to process modelling in the drinking water treatment industry. In developing models of multiple-input, multiple-output (MIMO) water treatment processes, a key consideration is the ability of the model to adapt to the dynamic nature of the process on a real-time basis. ANN models can handle non-linear

relationships, and can provide predictions of output variables in real-time in response to simultaneous and independent or dependent fluctuations in the values of model input variables. With respect to data processing, the type of relationship between the input and output data is determined purely from the information presented, with no presumptions from the ANN (Harvey and Harvey 1998). In addition, the ANN technique is fault-tolerant both in model development and in subsequent applications; discontinuities in the data, different levels of data precision and noise, and data scatter are easily accommodated (Foody and Aurora 1997). The technique is also extremely fast and flexible; advances in computing power have minimized the time required to develop models as well as the time required to re-train models and to incorporate new data. As well, ANN process models can be developed without quantifying the micro-scale interactions that occur. In drinking water treatment, such interactions are often poorly quantified, making it impossible to develop useful mechanistic process models. Also noteworthy is the fact that since ANN models are developed using full-scale operational data, the scale-up concerns commonly associated with bench-scale and pilot-scale empirical models are eliminated. Finally, ANNs do not require complicated programming, logical inference schemes, or the development of complex algorithms to build a successful model. Many user-friendly ANN software packages exist, offering the user a myriad of modelling options and allowing the user to customize the modelling process to suit his or her knowledge of modelling heuristics.

With respect to the disadvantages of the ANN modelling technique, many researchers consider the developed models to be "black-box" models, as ANNs do not yield explicit

mathematical formulae (Harvey and Harvey 1998). In addition, little is known about the applicability of the models to data that lie outside the domain on which the models were trained. As well, no set protocol for developing ANN models exists; each modeller may incorporate different modelling techniques. Finally, the ANN technique is data intensive and is best suited to problems where large data sets exist (Zhang and Stanley 1997). Current research efforts are aimed at eliminating or reducing the effects of these disadvantages in order to encourage the more widespread use of the ANN technique.

1.1.2. Existing ANN Models in the Water Treatment Industry

The drinking water treatment industry is an ideal candidate for ANN modelling and control applications. Both the quantity and quality of archived data in the industry continue to increase. Utilities archive operational data in order to better understand unit operations and to comply with government regulations. These same regulatory pressures, along with improvements in water quality and operational data measurement technologies, result in more robust data sets. In recent years, water treatment utilities have come to realize the potential of the ANN modelling technique, resulting in a number of published model applications in water quality, treatment, and distribution. With respect to water quality, models have been developed for trihalomethane (THM) formation and speciation (Hutton, Sandhu and, Chung 1996), source water salinity forecasting (DeSilets *et al.* 1992), and raw water colour forecasting (Zhang and Stanley 1997). Process models have been developed for coagulant and coagulant-aid dose forecasting in coagulation (Mirsepassi, Cathers, and Dharmappa 1995), and turbidity and colour removal through

enhanced coagulation (Baxter, Stanley, and Zhang 1999; Stanley *et al.* 2000). Finally, for distribution systems, models have been developed for the prediction of residual chlorine in the distribution system (Rodriguez *et al.* 1997), and for predicting water main breaks (Sacluti, Stanley, and Zhang 1999).

1.1.3. Integration of ANN Models for Process Optimization and Control

In water treatment plants, the completed models can be integrated into the supervisory control and data acquisition (SCADA) systems or stand-alone computers through graphical user interfaces (GUIs). These ANN interfaces serve as essential links between the ANN models, process data, the plant SCADA system, and end-users. Interfaces can be developed for both offline and online applications. In the former, the user manually enters model input data into target locations within the interface and model predictions are subsequently generated. In the latter, the interface receives model input data in real-time from the main SCADA computer. The model input data are processed through the ANN model and a model-predicted output value is returned to the interface.

With respect to process control applications, ANNs can be incorporated into an internal model control (IMC) scheme in either a direct or an indirect method (Psichogios and Ungar 1991). In the indirect method, the model is a process model trained to predict the output of the process, as previously discussed. As such, given the values of the process inputs and manipulated variables, the model predicts the expected value of the process output. In the direct method, a process-inverse model is trained to predict the value of a

manipulated variable required to reach a target value of the process output. As such, given the values of the process inputs, the values of all but one of the process control variables, and a desired value of the process output variable, the model predicts the optimal value of a process variable. Both the direct and indirect methods, alone or in combination, can be used to develop reliable automated process control systems.

1.2. RESEARCH NEEDS

In spite of the numerous existing research applications of the ANN technology to water treatment processes, several key research needs must be addressed in order to ensure the widespread application of ANNs in the water treatment industry. First and foremost, a systematic approach to developing ANN models in the water treatment industry currently does not exist. Without a model development protocol, utilities must rely on ad-hoc methodologies and chance to develop ANN models. In addition, the scope of existing applications has generally been limited to large and highly variable data sets that are well suited to ANN modelling. Little is known about the ability of ANN models to train on the small data sets traditionally encountered at new facilities or those with little process instrumentation. Likewise, the ability of ANN models to extract key relationships between input and output variables where one or more variables show little variation is not well understood. In addition, the boundaries of ANN model predictions and model extrapolation capabilities need to be fully assessed before full-scale process control applications can be developed. Finally, the application of ANNs to data analysis and in advanced process control in the water treatment industry remains to be thoroughly

studied. While theoretical discussions of such applications exist, there are few published accounts of successful model use in these areas.

1.3. OBJECTIVES

The overall goal of the research program was to develop artificial neural network (ANN) models for various process variables at water treatment facilities, and subsequently apply the models in data analysis and process optimization. This goal was achieved through the completion of the following objectives:

1. Create a systematic protocol for developing and evaluating ANN models in the drinking water treatment industry
2. Develop and evaluate ANN process models of filter effluent particulate concentrations at two facilities owned and operated by the Metropolitan Water District of Southern California (MWD)
3. Apply the models developed for MWD facilities in virtual laboratory experimentation and scenario analysis
4. Determine ANN model extrapolation capabilities in order to define ANN model prediction boundaries
5. Apply the ANN technology to the analysis of pilot-scale data
6. Develop and implement the water treatment industry's first-known ANN-based advanced process control system

In successfully completing these objectives, many of the research needs previously identified were addressed. A specific discussion of each objective, along with the needs addressed follows.

1.3.1. Creation of a Model Development Protocol

The first objective of the research program involved the creation of a systematic methodology for developing ANN models in the drinking water treatment industry. While the number of ANN-based publications in the industry has increased in recent years, guidelines and protocols for model development and implementation simply did not exist. A detailed framework for developing ANN models was developed and issues surrounding data collection and analysis, software selection and implementation, and model development and evaluation, were all thoroughly examined.

1.3.2. Modelling at MWD Facilities

In this component of the research program, models that predict the amount of particulate matter in filter effluent at Metropolitan Water District of Southern California facilities were developed. More specifically, models that predict filter effluent turbidity and filter effluent particle counts were developed for both the F.E. Weymouth Filtration Plant and the Oxidation Demonstration Project (ODP) Plant, both located in La Verne, California. While ANN models for many water treatment unit processes have already been developed, this objective contributes new knowledge to the field of water treatment

process modelling. Many of the existing ANN models in the industry have been developed using large, extremely variable data sets. The source water for MWD facilities is of good quality with little daily variation. As well, the ODP plant data set contained fewer than 80 data records, small by ANN modelling standards. In combination, these two factors served to challenge the ANN modelling process and, through model development, new insight into data requirements for successful modelling was gained.

1.3.3. Virtual Laboratory Experimentation and Scenario Analysis

Trained ANN models can be used both offline and online to provide operational information and assistance to plant operators. This objective involved the application of ANN models developed for MWD facilities in both virtual laboratory experimentation and scenario analysis. The virtual laboratory applications enable plant operators to determine optimal process conditions for a given raw water quality. In comparison to bench-scale experiments, such as jar tests, the virtual laboratory method is less time-consuming and labor intensive, and is not subject to scale-up concerns. Virtual experiments can also be performed to gain new insight into particle removal. Any of the model input variables can be varied, alone or in combination, in order to determine the effects of the changes on the filter effluent variables. When used in scenario analysis, ANN models allow the user to determine the cause of operational failures, as well as identify the impacts of proposed control actions on future operations. In combination these tools can greatly enhance the operational efficiency of water treatment facilities.

1.3.4. Defining Model Boundaries

One of the main concerns surrounding the application of ANN models in process control applications is the determination of model prediction boundaries. ANNs interpolate well within their training domain; accurate predictions can be made as long as the values of each of the input variables falls within the range of values on which the models were developed. Little is known however, about the ability of ANN models to provide accurate predictions for data outside this training domain. ANN models were developed using truncated data sets derived from complete full-scale operational data at EPCOR Water Services facilities. By applying the trained models to data outside the training domain, an assessment of model prediction boundaries was performed.

1.3.5. Analysis of Pilot-scale Data

Pilot-scale testing continues to be an integral part of implementing new drinking water treatment processes and process modifications. The primary goal of pilot testing is to extract the maximum amount of unbiased information regarding the factors affecting a particular process from as few observations as possible. Most experimental design techniques allow for the determination of the effects of many controlled or fixed factors on a single process output variable. Unfortunately, the effects of changes in the values of uncontrolled or random factors, which can vary considerably over the course of an experimental program, are not easily accommodated. This objective involved the application of the ANN technology to the analysis of pilot-scale data generated from two

separate facilities. This alternative methodology does not require special consideration of random factors and will undoubtedly find future use in the water industry.

1.3.6. Advanced Process Control

It has been theorized that ANNs can be incorporated as the control logic in advanced process control systems in the drinking water treatment industry. A framework for applying an ANN-based internal model control scheme in coagulant dosing was proposed by Zhang and Stanley (1999), however, the system was never implemented in real-time operations. This objective involved the development of a series of ANN models of the coagulation process at the EPCOR Water Services pilot plant, located in Edmonton, Alberta. The models were integrated into the plant supervisory control and data acquisition (SCADA) system to enable real-time control. The integrated models were processed through a custom-designed interface that iteratively determined the most efficient operational conditions required to maintain clarifier effluent quality within user-defined specifications. The results of this study suggest that full-scale ANN-based automation may be possible in the near future, resulting in the potential to reduce operating costs while improving finished water quality.

1.4. ORGANIZATION

In order to preserve the diversity of the models and applications developed in meeting the research objectives, a paper format has been employed in preparing this document.

Chapter 2 presents further background information on the ANN technology, as well as a systematic protocol for ANN model development and evaluation in the drinking water treatment industry. In Chapter 3, the results of ANN model development and evaluation at MWD facilities, as well as a number of developed model applications, are presented. Chapter 4 focuses on the evaluation of model prediction boundaries under different modelling conditions. In Chapter 5, the application of the ANN technology to the analysis of pilot-scale data is discussed. Chapter 6 presents the development, implementation, and evaluation of the drinking water treatment industry's first-known ANN-based advanced process control system. Finally, an overall assessment of the research program, as well as recommendations for future study, is presented in Chapter 7.

1.5. REFERENCES

Baxter, C.W., Stanley, S.J., and Zhang, Q. (1999) Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *Journal of Water Services Research and Technology – AQUA*. **48**(4):129-136.

DeSilets, L., Golden, B., Wang, Q., and Kumar, R. 1992. Predicting salinity in the Chesapeake Bay using backpropagation. *Computers and Operations Research*, **19**(3-4): 277-285.

Foody, G.M. and Aurora, M.K. 1997. An evaluation of some factors affecting the accuracy of classification. *International Journal of Remote Sensing*. **18**(4): 799-810.

Harvey, S. and Harvey, R. 1998. An introduction to artificial intelligence. *Appita Journal.* **51**(1): 20-24.

Hutton, P.H., Sandhu, N., and Chung, F.I. 1996. Predicting THM formation with artificial neural networks.[CD-ROM] In *Proceedings of the North American Water and Environment Conference*, American Society of Civil Engineers, Anaheim, CA. 7 p.

Jain, A. K., Mao, J.C., and Mohiuddin, K.M. 1996. Artificial neural networks: a tutorial. *Computer*. **29**(3): 31-44.

Mirsepassi, A., Cathers, B., and Dharmappa, H.B. 1995. Application of artificial neural networks to the real time operation of water treatment plants. *IEEE International Conference on Neural Networks: Proceedings.* Institute of Electrical and Electronics Engineers, Perth, Australia, pp. 516-521.

Psichogios, D.C. and Ungar, L.H. 1991. Direct and indirect model based control using artificial neural networks. *Industrial and Engineering Chemistry Research.* **30**: 2564-2573.

Rodriguez, M.J., West, J.R., Powell, J., and Serodes, J.B. 1997. Application of two approaches to model chlorine residuals in Severn Trent Water LTD. (STW) distribution systems. *Water Science and Technology.* **36**(5): 317-324.

Sacluti F., Stanley, S.J., and Zhang, Q. 1999. Use of artificial neural networks to predict water distribution pipe breaks. In *Proceedings of the 51st Annual Conference of the Western Canada Water and Wastewater Association.* Saskatoon, SK: WCWWA. 12 p.

Stanley, S.J., Baxter, C.W., Zhang, Q., and Shariff, R. 2000. *Process Modelling and Control of Enhanced Coagulation.* Denver, CO: AWWARF and AWWA. 167 p.

Zhang, Q. and Stanley, S.J. 1999. Real-time water treatment process control with artificial neural networks. *Journal of Environmental Engineering.* **125**(2):153-160.

Zhang, Q. and Stanley, S.J. 1997. Forecasting raw-water quality variables for the North Saskatchewan River by neural network modelling. *Water Research.* **31**(9): 2340-2350.

2. DEVELOPING ARTIFICIAL NEURAL NETWORK MODELS OF DRINKING WATER TREATMENT PROCESSES: A GUIDE FOR UTILITIES *

2.1. INTRODUCTION

Unit processes in the drinking water treatment industry are typically complex, involving many biological, physical, and chemical phenomena. While the numerous variables that directly influence the performance of each process are known to plant operators and researchers, the interactions and relationships between process inputs and outputs are often poorly understood and cannot be easily quantified. Process models, where they exist, are often site-specific and developed using the results of bench-scale and pilot-scale experiments where many of the process variables are held constant. Such models are unable to cope with simultaneous fluctuations in more than one or two key variables.

The inability of water treatment utilities to quantify process interactions and relationships can cause great difficulty for water treatment process control. If each unit process is considered to be a multiple-input multiple-output (MIMO) non-linear optimization problem, effective control can only be achieved through strategies that can accommodate the fluctuations of multiple input and output variables on a real-time basis. In the absence of such strategies, utilities resort to using bench-scale tests, which provide only a snapshot of the process conditions, in order to achieve process control.

* A version of this chapter has been submitted for publication. Baxter, C.W., Stanley, S.J., Zhang, Q., and Smith, D.W. Developing artificial neural network models of water treatment processes: a guide for utilities. *Journal of Environmental Engineering and Science*. Submitted 08/2001. 38p.

As utilities strive to meet more stringent customer-imposed and regulatory demands on finished water quality, there is an urgent need for new tools to improve process knowledge and provide effective alternatives to current process control methodologies. One such tool is artificial neural network (ANN) modelling, a robust technique that allows for the development of multiple-variable, non-linear models that can be integrated into real-time process control strategies.

Presented are guidelines for the development of ANN models for drinking water treatment unit processes. More specifically, issues surrounding the assessment of modelling needs, data collection and analysis, model protocol implementation, and model evaluation are discussed.

2.2. THE ARTIFICIAL NEURAL NETWORK TECHNOLOGY

ANNs are categorized as an artificial intelligence modelling technique due to their ability to recognize patterns and relationships in historical data and subsequently make inferences concerning new data. ANNs can be used for two broad categories of problems: data classification and variable prediction. For data classification problems, the ANN uses a specified algorithm to analyze data cases or patterns for similarities and then separates them into a pre-defined number of classes. Credit agencies, for example, often use ANN models to separate a potential list of clients into classes according to their level of credit risk or buying patterns. For variable prediction problems, the ANN learns to accurately predict the value of an output variable given sufficient input variable

information. The main applications of the ANN technique in the water treatment industry are in the development of water quality and process models and model-based process-control and automation tools. These applications can be categorized as variable prediction problems, to which the ensuing discussion is dedicated.

2.2.1. Types of ANN Models

Each drinking water treatment unit process can be considered a non-linear MIMO process, as previously discussed. The process inputs include influent water quality variables, such as pH and turbidity, as well as operational variables, such as chemical dosing levels and flow rate. The process outputs are the effluent water quality variables. Two different forms of ANN models of drinking water treatment process can be developed: process models and inverse process models. In the former, the model predicts the value of one or more process outputs, given the values of the process input variables. An example of this type of model is the prediction of clarifier effluent turbidity using influent water quality variables and operational variables. In the latter, the ANN model predicts the value of one or more process inputs, given the values of the remaining process inputs and process output(s). This type of model is often used to predict the value of an operational variable required to reach a target effluent quality. An example of this type of model is the prediction of alum dose required to maintain a desired value for clarifier effluent turbidity.

2.2.2. Key Components of ANN Models

There are several key components of ANN models that are collectively referred to as the ANN architecture. Processing units or neurons perform primitive operations such as scaling data, summing weighted inputs, and amplifying or thresholding sums. Neurons are organized into layers with each layer performing a specific function. The input layer serves as an interface between the input variable data and the ANN model. Most models also contain one or two hidden layers, although more are possible. These layers perform most of the iterative calculations within the network. The output layer serves as the interface between the ANN model and the end-user, transforming model information into an ANN-predicted value of the output variable(s). Each neuron is connected to every neuron in adjacent layers by weights; links that represent the 'strength' of connection between neurons. Each ANN model has a propagation rule that defines how the weights connected to a neuron are combined to produce a net input. The propagation rule is generally a simple summation of the weights. As discussed, the input layer serves as an interface between input variable data and the model. In this layer, a scaling function is used to scale data from their numeric range into a range that the network deals with efficiently, typically 0 to 1. The hidden and output layers contain an activation function that defines how the net input received by a neuron is combined with its current state of activation to produce a new state of activation. The most common activation function used in process modelling is the logistic activation function, although Gaussian, linear, and other functions can also be applied. Finally, each network has a learning rule that defines how the weights are modified in order to minimize prediction error. As will be

discussed, the backpropagation algorithm is the most common learning rule employed in process modelling. A schematic diagram and related discussion surrounding key components of ANN models is presented by Baxter *et al.* (2001a).

2.2.3. ANN Learning

In developing ANN process models, a supervised learning paradigm is employed. In supervised learning, historical data patterns, consisting of values for each of the model input and output variables, are used to train the ANN. The goal of supervised learning is to minimize the error between the model-predicted value and the actual value of the output variable(s). The error minimization takes place by modifying the weights between neurons according to a learning rule. The backpropagation ANN training algorithm is employed in the development of all models discussed in subsequent chapters. This algorithm increases or decreases the value of the weights in order to minimize the squared difference between the network-predicted value and the actual value of the output variable(s), summed over all of the data patterns. Training proceeds through repeated presentation of data patterns to the ANN and subsequent weight modification until the prediction error is sufficiently small, as defined by the user, or until a maximum number of iterations has been reached. A more detailed description of the backpropagation process is presented by Baxter, Stanley, and Zhang (1999).

The backpropagation algorithm is based on the generalized delta rule for learning. This rule can be summarized by the following three equations (Rumelhart, Hinton, and Williams 1986).

$$\Delta_p w_{ji} = \eta \delta_{pj} o_{pi} \tag{2.1}$$

$$\delta_{pj} = (t_{pj} - o_{pj}) f_j'(net_{pj}) \tag{2.2}$$

$$\delta_{pj} = f_j'(net_{pj}) \sum_k \delta_{pk} w_{kj} \tag{2.3}$$

In Equation 2.1, the change in the weight (Δw) from the j^{th} to the i^{th} unit following the presentation of pattern p is proportional to the product of an error signal, δ , available to the unit receiving input along that line and the output (o) of the unit sending activation along that line. The symbol η represents the learning rate of the system and has a value between 0 and 1. Equations 2.2 and 2.3 represent the error signal, the determination of which is a recursive process starting with the output unit. If a unit is an output unit, its error signal is given by the second equation where t_{pj} is the target input for the j^{th} component of the output pattern for pattern p, o_{pj} is j^{th} element of the actual output pattern produced by the presentation of input pattern p, and $f_j'(net_{pj})$ is the derivative of the semi-linear activation function which maps the total input to the unit to an output value. Finally, the error signal for hidden units for which there is no specified target is determined recursively in terms of the error signals (o_{pk}) of the units to which it directly connects and the weights (w_{kj}) of those connections.

2.2.4. Advantages of ANN Modelling

The ANN modelling technique holds several advantages over mechanistic modelling that make it particularly suitable to process modelling in the drinking water treatment industry. In developing models of MIMO water treatment processes, a key consideration is the ability of the model to adapt to the dynamic nature of the process on a real-time basis. ANN models can handle non-linear relationships, and can provide predictions of output variables in real-time in response to simultaneous and independent fluctuations of the values of model input variables. ANN models are also fault-tolerant in model building and in end-use. Data patterns where the value of one or more of the model inputs are missing can be incorporated into model building if necessary, as the modelling software can replace missing values with average values for the inputs involved. Obviously, model predictions will be more accurate if only complete data patterns are used. Similarly, when completed models are applied to new data patterns where values are missing, the value of model outputs can still be predicted. This feature is particularly useful in process control applications where input data is fed to models in real-time by instruments that are subject to periodic failure. Finally, ANNs do not require complicated programming, logical inference schemes, or the development of complex algorithms to build a successful model. Several user-friendly ANN software packages exist, offering the user a myriad of modelling options and allowing the user to customize the modelling process to suit his or her knowledge of modelling heuristics.

2.2.5. Challenges of ANN Modelling

Some aspects of the ANN modelling technique may present challenges to drinking water treatment utilities that wish to develop successful process models. ANN models can only be developed where sufficient historical data for each of the process variables exists. While data requirements will be addressed further on, it is important to note at this point that not all utilities have accurately detailed historical records of influent water quality, process control actions, and treated water quality. As more utilities realize that archiving such data is important for increasing process knowledge and improving process efficiency, this challenge will become less influential. Perhaps the greatest challenge that must be overcome is the perception of ANN models as black-box models that cannot be understood by the end-user. This perception is fed by a host of new ANN software packages that use proprietary algorithms to develop quick and easy models with minimal user interference. From a public health standpoint, utility managers are rightfully concerned about trusting plant operations to such a model. By using less sophisticated software packages that feature well-understood training algorithms, the effects of this challenge can be minimized.

2.2.6. ANN Applications

Once developed, ANN models can be applied in a number of off-line and on-line process assessment and control applications that involve varying degrees of sophistication. The least complex application of the ANN technique in the drinking water treatment industry

involves the use of ANN models for process assessment and data analysis. In such applications, ANN models can be used to identify and assess difficulties in plant operations and suggest potential remedies. The technique can also assist operators in determining the effects of typical operating conditions on a newly measured treated water variable. Process assessment and data analysis applications generally involve the development of an ANN process model. Utilities that have a more extensive historical database, consisting of a wide variety of reliable water quality and operational data can use the ANN technique to develop a number of offline tools to assist operators in daily plant operations. ANN models can be successfully applied in scenario analysis, operator training, and virtual laboratory applications. When integrated into the plant SCADA system, these tools can also be executed on-line using real-time data. The most sophisticated level of ANN applications involves the use of trained models in real-time advanced process control, whereby a trained ANN model is used as the control logic in an automated control loop. A more thorough discussion of existing and potential applications of the ANN technology in the water treatment industry is presented by Baxter *et al.* (2001a; 2001b).

2.2.7. Existing ANN Models

In recent years, water treatment utilities have come to realize the potential of the ANN modelling techniques, resulting in a number of published model applications in water quality and demand forecasting, treatment, and distribution. With respect to water quality and demand forecasting, models have been developed for trihalomethane (THM)

formation and speciation (Hutton, Sandhu, and Chung 1996), source water salinity forecasting (DeSilets *et al.* 1992), raw water colour forecasting (Zhang and Stanley 1997), and water demand forecasting (Baxter *et al.* 2001a). Process models have been developed for alum and polymer dose forecasting in coagulation (Mirsepassi, Cather, and Dharmappa 1995), lime dose and hardness in softening (Baxter *et al.* 2001a), and turbidity and colour removal through enhanced coagulation and filtration (Stanley *et al.* 2000). The control of coagulation has been demonstrated by Baxter *et al.* (accepted 06/2001). Finally, for distribution systems, models have been developed for the prediction of residual chlorine (Rodriguez *et al.* 1997) and the prediction of distribution system pipe breaks (Sacluti, Stanley, and Zhang 1999).

2.3. BUILDING ANN MODELS OF DRINKING WATER TREATMENT PROCESSES

In order to build effective ANN models of drinking water treatment processes, a sequential methodology consisting of four key stages is proposed. The relationship between each of the stages is depicted in Figure 2.1, while a detailed description of each stage is presented in the ensuing sections.

2.3.1. Needs and Suitability Assessment

The first stage of successful ANN model development involves an assessment of the utility's needs regarding the model and its applications, as well as the suitability of the

ANN technique for the problem at hand. With regards to the latter, specific data, software, hardware, and personnel requirements must be met to take advantage of the ANN technology. The key requirement of the ANN modelling approach is the availability of relevant data to describe the process being modelled. Data must be available in a useable digital format for each of the process input and output variables. The data used in model development must be representative of plant operations, spanning the range of operating conditions that may be encountered during both routine operations and process upset conditions. Model development also requires the use of appropriate ANN software. Many commercial software packages are currently available; site-specific restrictions on operating systems and data format, as well as desired options will dictate the best software choice for each utility. As a general rule, software packages that use only proprietary algorithms should be avoided in favour of those that have a high degree of user input in model-building. With respect to hardware requirements, recent advances in computing technology have made even the most modest new home computing system capable of performing the modelling calculations. To ensure optimal performance however, a 500 MHz processor and 128 M of RAM should be considered to be the minimum system requirements. With respect to personnel requirements, the model developer should have expert knowledge of the process being modelled and should understand basic modelling heuristics. As such, an operations engineer with some training in ANN model development would be a suitable candidate to develop process models. Finally, it is important to recognize that the ANN technology is most suitable for modelling of complex non-linear processes where the interactions between process inputs and outputs are poorly understood. Where processes are well described by existing

mechanistic models or site-specific empirical models, the ANN technology may not offer improved modelling results.

With regard to the utility's needs, the ANN model development process can be tailored to ensure that the intended model applications are successful. Once developed, ANN models can be applied in a number of off-line and on-line process assessment and control applications, as previously discussed. Where the intended application of the model is in the analysis of process data or assessment of process performance, a process model should be developed. In some real-time process control applications however, process inverse models are required. Furthermore, the level of accuracy of model predictions increases with the sophistication of the model application. In real-time chemical dosing control, for example, the level of tolerance for error is far less than that for a simulation tool used to train plant operators. By ensuring that appropriate models are developed from the beginning, inefficiencies in the modelling process can be minimized.

2.3.2. Data Collection and Analysis

In order to develop successful ANN models of drinking water treatment processes, careful attention must be paid to the details of data collection and analysis. In collecting data, several factors need to be considered. First, the availability of the data must be ascertained. For data availability, the variables for which historical data exist, the time frame of historical measurements, and the frequency of data measurement must all be determined. The format and reliability of the data are also key considerations in data

collection. Historical data can originate from grab-samples or real-time measurements, and measurements can be discrete or aggregated from a number of samples. The reliability of the available data should be ascertained through an examination of quality assurance and quality control protocols. Finally, it is of considerable importance to note any process changes that may have been implemented during the time frame for which data are available.

With the above considerations in mind, it is possible to delineate a number of guidelines for selecting data to be used in ANN process modelling. First and foremost, data for each of the variables known or suspected to affect the process being modelled must be available. The quantity of data required to develop a model is site specific, and is affected by seasonal fluctuations in influent water quality and the frequency of process upsets. As such, it is important to ensure that the data are fully representative of the range of conditions that can be expected during periods of routine and upset operations. As a general guideline, at least one full cycle of data must be available in order to ensure a representative data set. In temperate areas with four distinct seasons for example, a data cycle would encompass a full year of historical data. With respect to the format of the data, the variability of the process as well as data availability will dictate whether to use hourly data, daily averages, or some other frequency for each of the variables. Successful process models can often be made using the daily average or some daily percentile value of each of the model variables. With respect to the effect of major process changes on data selection, data collected prior to major process changes should not be incorporated if they are believed to differ significantly from current process data. This guideline is

necessary to ensure that the data are representative of current process conditions. Finally, in order to maintain the integrity of the data set, appropriate quality assurance and quality control protocols for the collection of each model variable must be in place.

Once an appropriate historical data set has been selected, it should be fully characterized and subjected to a comprehensive statistical analysis. Data characterization involves a qualitative assessment of hourly, daily, and seasonal trends of each potential model variable. The statistical analysis involves the determination of measures of central tendency, measures of variation, and a percentile analysis, as well as the identification of outliers, erroneous entries, and non-entries for each data variable. In combination, the data characterization and statistical analysis help to identify the boundaries of the study domain as well as potential deficiencies in the data set.

2.3.3. Application of the Model Building Protocol

Unfortunately, there is no widely accepted best method of developing ANN models. When all of the possible options in building the ANN model architecture are considered, an almost infinite number of distinct architectures are possible. As such, each model developer may use a different protocol to reduce the number of architectures that are evaluated. What follows is a five-step protocol that the author has found to be useful in developing drinking water treatment ANN process models, represented schematically in Figure 2.1.

2.3.3.1. Selection of Model Inputs and Outputs

The first step involved in the selection of model inputs and outputs is the selection of the model output(s). The output(s) are selected on the basis of operational needs, relevant literature, and data availability. As previously discussed, drinking water treatment processes can be considered MIMO processes. As such, process models can have multiple output variables. Since ANN models train by minimizing the error between the predicted and actual values of model output variables, however, the technique yields better results when a single output is modelled. Where it is desirable to model more than one process output variable, separate models should be developed for each output to reduce overall prediction errors. Once the model output has been selected, model inputs are selected from the available variables. Input selection is based on the existence of a known or suspected relationship with the output variable, relevant literature, and data availability. Initially, it is better to include all applicable variables, as those found to be redundant or not important could be removed in subsequent trials. Where possible, the use of lag values, previous values of a variable in a time series, as input variables should be avoided. The inclusion of multiple lag variables can result in developing a model that performs time-series forecasting as opposed to one that maps input and output relationships. As well, the resulting model cannot be easily transferred to real-time applications where the data sampling frequency differs significantly from the lag period.

2.3.3.2. Selection and Organization of Data Patterns

Once the model input and output variables have been identified, the modelling data sets can be constructed. Each data pattern or record should initially be examined for erroneous entries, outliers, and blank entries. Outlier detection involves a high degree of subjectivity. All values that are outside a range of ± 2 standard deviations from the mean of a variable may be excluded from the data set, for example. Alternatively, scatter plots of each variable can be used to detect outlier values. Data patterns that contain questionable data should be removed, and a record of removed patterns kept for future reference and analysis. The remaining data patterns must be divided into three data sets: the training set, the test set, and the production set. The training set is the largest set and is used to train the model, as previously discussed. The test set serves as a semi-independent check on the progress of ANN learning. Without the test set, the model would simply memorize the interactions present in each of the training patterns and would not be able to provide accurate predictions on data from outside the training set. Most ANN software packages periodically process the test set through the model during training to ensure that memorization does not occur. The production set is used as an independent validation of the model following training. The trained model is applied to the production set data patterns, to which the model has not been exposed, and an assessment of the accuracy of prediction is made. The data patterns are divided among the training, test, and production sets in a predetermined ratio; a 3:1:1 ratio has proven to be effective for many process models. Sorting the data according to the value of the output variable and then assigning the patterns to the data sets in order according to the

ratio, has proven to be an effective means of ensuring that the three data sets are similar with respect to the mean and variance of the output variable and that each data set is fully representative of the study domain. Statistical differences between the data sets can be detected through the use of ANOVA and other statistical measures.

2.3.3.3. Determination of Architecture Characteristics

The determination of architecture characteristics is the step where the model architecture is actually built. Depending on the ANN software being used, many different architecture factors can be selected and varied by the user. Some of the most common factors include: base architecture type, number of hidden layers, number of hidden layer neurons, type of scaling function, type of activation functions, initial range of weights between neurons, and the type of learning rule (Table 2.1). The characteristics of each of these factors are discussed in great detail by Baxter (1998). Evaluating the best values for each of the many factors can take a considerable amount of time. On the basis of past modelling experiences, successful process models can be developed using a multilayer perceptron network with a single hidden layer, a linear scaling function in the input layer, logistic activation functions in the hidden and output layers, random initial weight values, and the backpropagation learning rule (Stanley *et al.* 2000). This configuration is commonly referred to as a standard three-layer multiplayer perceptron ANN architecture. When using this standard architecture, the only major factor that needs to be experimentally determined is the number of hidden layer neurons. By creating a series of models that differ only in the number of hidden layer neurons and recording the statistical measures

of prediction error for each, the best models can be identified. While the prediction error can be assessed by a number of different statistical methods, the best models will generally have a low mean absolute error and high coefficient of multiple determination (R^2) when applied to the production set. At this stage, it is not uncommon to have multiple candidate models that offer similar prediction performance on the production set. The final model is selected from among these candidates in future steps.

2.3.3.4. Evaluation of Model Stability (Cross-Validation)

In order to ensure that the prediction performance of the candidate models is independent of the manner in which the data patterns were separated into the three data sets, the patterns in the three sets are redistributed. A simple way to achieve pattern redistribution without affecting the 3:1:1 ratio is to move the first data pattern to the end of the data set after sorting in order of the value of the output variable and prior to assigning the patterns to one of the three modelling data sets. The candidate models are re-trained on the new data sets and the results are compared to those of the candidate model. A significant increase in prediction error on the new production set is an indication of model instability. The best candidate models will have similar prediction errors when the data sets are redistributed. If model instability is detected, the data should be re-sorted into the three modelling data sets using the advanced methodology discussed in Section 2.4.

2.3.3.5. Model Fine-tuning

In determining the architecture characteristics, the number of hidden layer neurons is the only major factor that is evaluated. In model fine-tuning, a number of software-specific variables, such as the type of scaling and activation functions and the type of learning rule, can be altered in an attempt to achieve modest decreases in prediction error on the production set. Model fine-tuning is typically user-specific and no protocols are known to exist. Some models do not improve during fine-tuning, and the improvement in those that do may not justify the time and effort required for this step. Where a maximum tolerable prediction error has been mandated for the process being modelled however, model fine-tuning can make the difference between acceptable and unacceptable prediction performance.

2.3.3.6. Repetition of Modelling Protocol Steps

The protocol presented thus has been successfully applied in the development of ANN models presented throughout the remainder of the document. With the exception of the minor software-specific modifications discussed in Chapter 3, no modifications to the protocol should be required to develop process models in the drinking water treatment industry. On occasion, new information concerning the process being modelled or the data used in modelling will be uncovered during the modelling process. It is not uncommon, for example, to discover that one or more variables initially included as model input parameters are redundant and can be removed without negatively impacting

model performance. In addition, previously undetected erroneous data patterns may be discovered at any point in the modeling process due to errors and omissions in data analysis and characterization. The sequential multi-step modelling approach presented in preceding sections and in Figure 2.1 can be augmented with recursive loops to accommodate the repetition of some of the modelling steps should the need arise.

2.3.4. Performance Evaluation

The model development protocol presented herein can lead to the development of several candidate models, each offering similar prediction capabilities. The best model is the one that meets the needs defined in the first stage while offering the smallest prediction errors. Prediction errors can be evaluated through a number of statistical and graphical methods. With respect to the former, both absolute and relative measures of error are often reported by the ANN modelling software, as are coefficients of correlation and determination. Absolute measures of error, such as mean absolute error (MAE) and maximum absolute error, allow the user to compare model performance to utility needs on an absolute scale. In building a model of turbidity removal through coagulation, for example, a utility may specify a maximum absolute error of 0.1 NTU for predictions made on the production set. As such, absolute error measurements allow immediate comparison to utility needs and targets. Relative measures of error, such as mean absolute percentage error (MAPE) can help to identify percent discrepancies between actual and predicted values and provide the user with a type of prediction error threshold. Finally, coefficients of correlation can be used to determine the strength of the relationship

between the actual and predicted values, while coefficients of determination explain how much of the variation in the predicted values is accounted for by the model.

Graphical analyses of model results provide visual confirmation of model prediction ability. By plotting either the predicted values of the model output variable or the absolute prediction errors, along with the actual historical values, across all data patterns, periods of acceptable and unacceptable model performance can be identified. Such plots are particularly useful for identifying events where the largest prediction errors occurred. Once these events have been identified, the user can determine the source of the prediction error; both erroneous data entries and incomplete model training can yield large prediction errors.

Where the completed model is to be used for online applications in real-time, the performance evaluation stage should also include an evaluation of model performance in real-time. Most models are built using historical data sets where values of model variables have been averaged on a daily basis or otherwise manipulated. When applied in real-time applications, the model must be able to effectively handle the discrepancies caused by differences in data collection frequency and methodology. Real-time model evaluations can be accomplished through integrating the model with the plant SCADA system and monitoring model predictions over a set time period. A detailed discussion of model integration is presented in Chapter 6. Alternatively, the model can be applied to a historical data set on a stand-alone computer for which the collection frequency simulates real-time. Since ANN models learn by mapping the underlying cause-effect relationships

between input and output variables, models developed on averaged or aggregate data can generally be transferred to real-time applications with little difficulty. Real-time model evaluations should be carried out to confirm that this observation holds for the modelling scenario being studied, before time and effort are spent on full-scale integration.

2.4. ADVANCED MODELLING

The stages of model development presented in the previous sections are sufficient for developing reliable models under most circumstances. Occasionally, utilities that have extremely complex raw water quality and operational profiles will benefit from the use of more advanced modelling techniques. While a complete discussion of advanced ANN modelling is beyond the scope of the current discussion, some points worthy of consideration are presented here.

At surface water treatment facilities that experience extreme seasonal variations in raw water quality, a more complex method of separating the modelling data into training, testing, and production sets may be required to ensure a representative distribution of data in each set. Typically, the raw water quality at these facilities can be categorized according to the values of one or more model input variables. In Edmonton, Alberta, Canada for example, there are seven different raw water categories that are observed sequentially throughout the year; data can be categorized by the values of raw water temperature, raw water turbidity, and raw water colour. The categorization of data patterns can be facilitated through the use of Kohonen ANNs. These networks

automatically separate all of the data patterns into a user-defined number of categories based on the values of one or more user-selected input variables (Baxter *et al.* 2001b). As such, data patterns with similar raw water quality attributes will be grouped together. Once the data patterns are categorized, representative training, testing, and production data sets can be built by separating the patterns in each category in a 3:1:1 ratio, as previously discussed.

Where the boundaries for the different raw water categories are well defined, and where plant operational strategies for these categories differ greatly, a separate ANN model can be developed for each category. The benefit of this strategy is that each model is trained on a more cohesive data domain, resulting in better model performance when compared to a single model for all categories combined. The multiple models can be neatly linked using a Kohonen network categorization system in order to facilitate real-time model applications. When input data are received by the system, the Kohonen network identifies the category to which the data belongs and the appropriate model can be applied. A detailed description of such a system is presented by Baxter *et al.* (2001b).

2.5. CONCLUSION

While the challenges involved in developing ANN process models may be prohibitive for some utilities, the current trend towards archiving process data, in combination with recent advances in computing technology, have made the technology appropriate for many utilities. The ANN modelling technique allows utilities to develop multiple-

variable, non-linear models of complex unit processes. As customer and regulatory demands on finished water quality continue to become more stringent, ANN models will continue to offer drinking water utilities a sophisticated process modelling and control alternative to conventional methodologies.

2.6. REFERENCES

Baxter, C.W. 1998. Full-Scale Artificial Neural Network Modelling of Enhanced Coagulation. Master's thesis. University of Alberta, Edmonton, AB. 151 p.

Baxter, C.W., Stanley, S.J., and Zhang, Q.. 1999. Development of a full-scale artificial neural network nodel for the removal of natural organic matter by enhanced coagulation. *Journal of Water Supply: Research & Technology – AQUA*, **48**(4):129-136.

Baxter, C.W., Zhang, Q., Stanley, S.J., Shariff, R., Tupas, R-R.T., and Stark, H.L. 2001a. Drinking water quality and treatment: the use of artificial neural networks. *Canadian Journal of Civil Engineering*, **28**(Suppl. 1): 26-35.

Baxter, C.W., Tupas, R-R.T., Zhang, Q., Shariff, R., Stanley, S.J., Coffey, B.M., and Graff, K.G. 2001b. *Artificial Intelligence Systems for Water Treatment Plant Optimization*. American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 141 p.

Baxter, C.W., Shariff, R., Stanley, S.J., Smith, D.W., Zhang, Q., and Saumer, E.D. Model-based advanced process control of coagulation. *Water Science and Technology*, 8 p. (accepted 06/2001)

DeSilets, L., Golden, B. Wang, Q., and Kumar, R. 1992. Predicting salinity in the Chesapeake Bay using backpropagation. *Computers and Operations Research*, **19**(3-4): 277-285.

Hutton, P.H., Sandhu, N., and Chung, F.I. 1996. Predicting THM formation with artificial neural networks. In *Proceedings of the North American Water and Environment Conference*, Anaheim, CA.: ASCE.

Mirsepassi, A., Cathers, B., and Dharmappa, H.B. 1995. Application of artificial neural networks to the real time operation of water treatment plants. *In IEEE International Conference on Neural Networks: Proceedings*. Perth, Australia: IEEE

Rodriguez, M. J., West, J.R., Powell, J., and Serodes, J.B. 1997. Application of two approaches to model chlorine residuals in Severn Trent Water LTD (STW) distribution systems. *Water Science and Technology*, **36**(5): 317-324.

Rumelhart, D.E., Hinton, G.E., and Williams J.R. 1986. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* D. E. Rumelhart, G. E. Hinton and R. J. Williams. Cambridge, MA: The MIT Press. 1: 318-362.

Sacluti F., Stanley, S.J., and Zhang, Q. 1999. Use of artificial neural networks to predict water distribution pipe breaks. In *Proceedings of the 51st Annual Conference of the Western Canada Water and Wastewater Association.* Saskatoon, SK: WCWWA. 12 p.

Stanley, S.J., Baxter, C.W., Zhang, Q., and Shariff, R. 2000. *Process Modelling and Control of Enhanced Coagulation.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO: 167 p.

Zhang, Q. and Stanley, S.J. 1997. Forecasting raw-water quality variables for the North Saskatchewan River by neural network modelling. *Water Research*, **31**(9): 2340-2350.

Table 2.1 Common factors and values in determining ANN architecture characteristics

Factor	Common Values
Base architecture type	Multilayer Perceptron, Radial Basis Function, General Regression, Probabilistic, Kohonen
Number of hidden layers	0 to 3
Number of hidden layer neurons	1 to >100 per hidden layer
Type of scaling function	Linear, logistic, tanh
Type of activation function	Logistic, linear, tanh, Gaussian
Initial range of weights	Any range between 0 and 1
Type of learning rule	Backpropagation, Conjugate Gradient Descent, Quasi-Newton, Levenberg-Marquardt, Kohonen

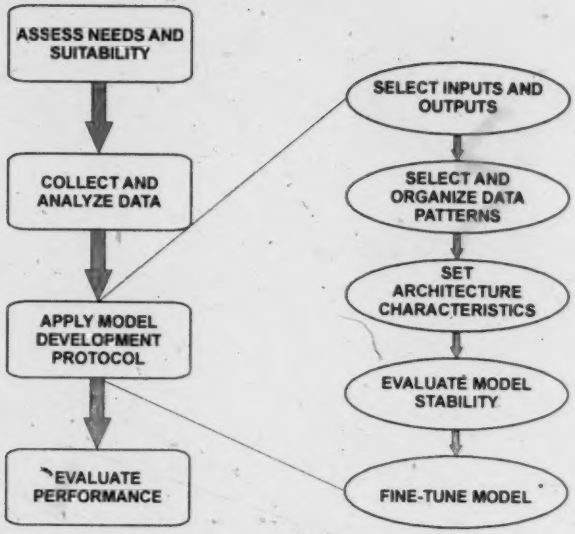


Figure 2.1 The main stages of developing an ANN process model

3. PROCESS MODELLING AND MODEL APPLICATIONS FOR METROPOLITAN WATER DISTRICT OF SOUTHERN CALIFORNIA FACILITIES*

3.1. INTRODUCTION

This chapter presents a detailed description of model development and results for two full-scale facilities that are owned and operated by the Metropolitan Water District of Southern California (MWD). Based on data availability and partner utility requirements, a number of models were developed to predict the values of key filter effluent variables at the Oxidation Demonstration Project (ODP) Plant and the F.E. Weymouth Filtration Plant. More specifically, models were developed for the prediction of filter effluent turbidity (NTU), and filter effluent particle counts ((counts > 2 μm)/mL) using historical data from each of the two facilities. Applications of the models in virtual experimentation and scenario analysis are also presented.

3.2. BACKGROUND INFORMATION

3.2.1. Oxidation Demonstration Project Plant

The ODP facility, located on the premises of the F.E. Weymouth Filtration Plant, began operation in 1992 in order to evaluate the use of ozone and PEROXONE (a combination of ozone and hydrogen peroxide) as primary disinfectants at full-scale MWD facilities.

---

* A version of this chapter has been published. Baxter, C.W., Tupas, R-R.T., Zhang, Q., Shariff, R., Stanley, S.J., Coffey, B.M., and Graff, K.G. 2001. *Artificial Intelligence Systems for Water Treatment Plant Optimization.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 141 p.

The facility has a capacity of 20.8 ML/d (5.5 MGD) and consists of two types of ozone generators, two feed gas systems, an external mix ozone contacting system, two ozone contactors, three types of ozone destruct systems, a hydrogen-peroxide injection system, and conventional coagulation and filtration unit processes. With respect to conventional unit processes, the ODP plant has the ability to feed sulfuric acid, sodium hydroxide, metal coagulant (alum, ferric chloride, or polyaluminum chloride), polymer, filter aid, chlorine, and ammonia. The ODP plant is operated as a stand-alone research facility, although treated water is currently pumped from the clearwell to the head works of the F.E. Weymouth Filtration Plant.

3.2.2. F.E. Weymouth Filtration Plant

The F.E. Weymouth Filtration Plant is a conventional treatment facility with a design flow of 1960 ML/d (518 MGD). The facility uses polymer-assisted alum clarification through 8 independent flocculation and sedimentation basin trains, followed by filtration through a bank of 48 dual-media filters. Free-chlorine is used as the primary disinfectant, and chloramines are used to provide a disinfectant residual in the distribution system. The facility receives Colorado River water (CRW) via MWD's 389 km (242 mile) long Colorado River Aqueduct, and State Project Water (SPW), originating in Northern California, via the 714 km (444 mile) long California Aqueduct.

3.3. ANN MODEL DEVELOPMENT

3.3.1. Methodology

3.3.1.1. Data Collection and Management

All of the data used in the development of the models were obtained from MWD, which owns and operates both the ODP and F.E. Weymouth facilities. In addition to the evaluation of alternative disinfectants, the ODP facility has been used to study biological filtration, filter air binding, enhanced coagulation, arsenic removal, and other filtration issues. Each of these studies required different plant configurations and had different data collection requirements. As such, while the plant had been in continuous operation for over 7 years at the time of the study, the data set was often fragmented or lacking continuity. A filtration study, conducted at the ODP Plant from July 1996 – December 1997, offers the most comprehensive data record at the facility. The majority of the water quality variables recorded during the study are based on grab sample water quality analyses, although filter effluent turbidity and particle counts were determined using online analyzers. A complete list of the water quality variables that were measured at the facility during this study is presented in Table 3.1, while the list of operational variables observed during the study is presented in Table 3.2. As many variables at MWD facilities continue to be measured in non-SI units, data in the later table, as well as in other tables

throughout the chapter, are presented in their measured units. Conversions factors are provided in table footnotes.

The data routinely collected at the F.E. Weymouth facility can be divided into three categories: water quality lab data, operational data, and online data. The water quality lab data consist of the values of various water quality variables, from samples collected at various locations throughout the plant and subsequently analyzed in the on-site water quality lab. A complete listing of the water quality lab data is presented in Table 3.3. The operational data consist of chemical feed rates, flow rates, and filtration data, as entered into operational logs by plant operators. A list of the available operational variables is presented in Table 3.4. The online data are collected from various online instruments by the plant's SCADA system, which has limited data storing capabilities. A list of the data variables collected by the online instruments at the plant is presented in Table 3.5. In addition to the variables listed in the three tables, limited particle count information was collected during an independent study at the facility in 1999. Particle count measurements were taken in the plant influent and combined filter effluent, as well as in the effluent of various filters.

3.3.1.2. Software

Data analysis and organization was accomplished using Microsoft Excel spreadsheets as well as the Statistica 5.5 from StatSoft of Tulsa, Oklahoma. Statistica Neural Networks (SNN) release 4.0E, also from StatSoft, was used for the development of most models.

This software includes features for the automation of certain repetitive tasks involved in ANN process modelling and is therefore more efficient than comparable modelling software. NeuroShell2 release 4.0, from Ward Systems Group of Frederick, Maryland, was used in the development of the particle count model for the F.E. Weymouth facility.

3.3.1.3. Model Development Protocol

As was discussed in Chapter 2, the development of ANN process models in the water treatment industry is best accomplished through the use of a five-step protocol: input/output selection, data pattern organization, architecture selection and modification, model stability evaluation, and model fine-tuning. This protocol ensures that models are developed in a systematic fashion and is applicable for a wide variety of modelling applications. In order to make the best use of SNN's automation features, however, some modifications to the general protocol were required. In SNN, architecture selection and modification, model fine-tuning, and initial comparison of candidate models are all accomplished using the Intelligent Problem Solver (IPS) module. The IPS module applies user-defined constraints, in combination with randomized initial weight values, to simultaneously evaluate a number of different architectures and identifies the best candidate models for further study. As such, the modified protocol employed in the development of ANN models for MWD facilities, where SNN was used as the modelling software, consisted only of four steps: input/output selection, data pattern organization, architecture selection and fine-tuning through the IPS module, and model stability evaluation. The IPS also employs the conjugate gradient descent learning rule in addition

to the backpropagation rule discussed in Chapter 2. Learning proceeds sequentially with a number of backpropagation training epochs followed by a number of conjugate gradient descent epochs. An epoch is defined as a period of learning in which each of the data patterns is presented to the network once. Conjugate gradient descent differs from backpropagation in that it determines the average gradient of the error surface across all patterns and updates all weights at the end of a training epoch, whereas backpropagation adjusts the weights after the presentation of each pattern.

3.4. RESULTS

3.4.1. Source Data Analysis

3.4.1.1. Oxidation Demonstration Project Plant.

The data set from the 1996-1997 filtration study consists of 125 individual data records. Each record presents water quality and operational data that have been averaged over the course of a single 24-hour filter cycle. As can be seen in Table 3.6, the raw water quality over the course of the study was relatively good. Influent turbidity exceeded 7.30 NTU less than 5 % of the time and never exceeded 9.60 NTU. As is noted in Table 3.6, some of the raw water quality variables, such as UV_{254} absorbance and alkalinity, were measured infrequently throughout the study, and all other variables were subject to frequent data collection errors that resulted in blank data entries. Some entries for dissolved oxygen concentration were found to exceed the saturation concentration due to

pressure changes in the system and algal growth in the raw water. With regards to seasonal variation of raw water quality, both temperature and turbidity tended towards higher values in the spring and summer months (Figure 3.1).

With respect to operational characteristics throughout the study, the plant was operated at a mean flow, reported with its standard deviation, of 14.5 ± 4.43 ML/d (3.83 ± 1.17 MGD), while mean doses of the ferric chloride coagulant and polymeric coagulant aid, reported with their standard deviations, were 4.17 ± 1.29 mg/L and 2.67 ± 0.72 mg/L, respectively (Table 3.7). With few exceptions, the filter aid feed was either disabled or was set to a dose of 0.01 mg/L. Ozone was used as the primary disinfectant in the study, with a mean dose, reported with its standard deviation, of 1.13 ± 0.34 mg/L.

As previously discussed, both filter effluent turbidity and filter effluent particle counts were monitored in real-time using online instruments. For the purposes of the filtration study, the data from the instruments were compiled and presented as the 50^{th} and 90^{th} percentile values over a 24-hour filter cycle (Table 3.7). With few exceptions throughout the study, the filter effluent turbidity was maintained below 0.10 NTU as evidenced by the data analysis for both the 50^{th} and 90^{th} percentile filter effluent turbidity values. Filter effluent particle counts demonstrated far more variation however, with a mean 50^{th} percentile value of 47 counts/mL and a standard deviation of 43 counts/mL.

3.4.1.2. F.E. Weymouth Filtration Plant.

The historical data set that was selected for use in the modelling of the F.E. Weymouth Filtration Plant spans from January 1997 to December 1999, the most current data available at the time of modelling. As can be seen in Table 3.8, the raw water is characterized by low turbidity, mirroring values observed at the ODP plant. Both the raw water temperature and turbidity are subject to seasonal variation, with higher values occurring in the summer months. As with the ODP plant, some dissolved oxygen concentration measurements exceeded the saturation concentration due to algal growth and pressure changes during treatment. As was previously mentioned, only limited particle count data, collected during an independent internal study, are available for the facility. As can be seen in Figure 3.2, the raw water particle counts do not appear to be subject to seasonal variation. With regards to raw water blend, the plant typically uses water from the Colorado River system exclusively during the winter months and blends with State Project Water (SPW) resources as they become available from spring to fall at a typical blend of 20 to 30% SPW.

The quality of the raw water is such that the facility uses low chemical doses to achieve treatment. The mean doses of alum and polymer used at the facility, reported with their standard deviations, were 4.2 ± 0.6 mg/L and 1.5 ± 0.2 mg/L, respectively (Table 3.9). Following filtration, water turbidity values were extremely stable and rarely exceed 0.1 NTU. Mean filter effluent particle counts were more variable, with a mean of 50th

percentile values (calculated on a 24-hour basis) equal to 13.8 counts/mL and a standard deviation of 25.7 counts/mL.

3.4.2. Model Development and Evaluation

3.4.2.1. Oxidation Demonstration Project Plant.

Models were developed for each of two separate but related indicators of filter effluent quality at the ODP plant: the 50th percentile value of filter effluent turbidity and the 50th percentile value of filter effluent particle counts. The 50th percentile, or median, values were chosen over the mean values as they are less affected by the existence of observations made at extreme ends of a variable's range.

Based on data availability and reliability, as well as recent literature and preliminary modelling results, seven model inputs were used in the development of each model. The inputs consisted of raw water quality variables and operational variables and are listed in Table 3.10. While the filtration study data set used in model development was the most complete data set available from the ODP facility, the frequency of measurement for each of the model variables varied considerably. The ANN modelling technique can incorporate data patterns that contain blank entries although, in practice, better models are obtained if such data patterns are not included in model development. As a result, data patterns containing blank entries, as well as those containing erroneous entries were removed. The resulting final modelling data sets consisted of only 80 and 75 data

patterns, of the original 125 patterns in the complete data set, for the turbidity and particle count models, respectively.

In any ANN process modelling, it is essential to ensure that the data used are representative of the process conditions observed through at least one cycle, as previously discussed. When small data sets are used in modelling, as was the case for the ODP models, extra care is needed in order to ensure that the data set is indeed representative. This can best be accomplished through an analysis of variance (ANOVA) between the original complete data set and the reduced modelling data set for each model variable. For almost all of the ANOVA analyses, the calculated value of the F statistic was less than the critical value for the statistic at the 0.05 level of significance, indicating no statistical difference between the variable values in the original complete data set and the reduced modelling data set. Four of the analyses, those for filter-aid dose and filtration rate for both models, yielded calculated values of the F statistic that exceeded the critical value. The statistically significant difference between the values of filter-aid dose between the original data set and each of the two modelling data sets can be explained by the nature of filter-aid dosing during the study from which the data were derived. As previously discussed, the filter-aid dosing system was either disabled or set to deliver a concentration of 0.01 mg/L. Both the modelling sets contain a higher proportion of patterns where the filter-aid dose was 0.00 mg/L when compared to the original data set. Each of the modelling sets therefore had a lower mean value and higher variance for filter-aid dose, resulting in a statistically significant difference when compared to the complete data set. In spite of this difference, each of the modelling sets contains data

patterns that are representative of each of the two possible filter-aid dosing conditions. As such, the validity of the modelling data sets is not affected by the statistically significant difference. With regards to ANOVA analyses for filtration rate, the statistically significant difference can be explained by the filtration rate having had a value of 0.00 gpm/ft^2 in at least 5% of the patterns in the complete data set (Table 3.7). These patterns were not included in the modelling data sets as the filtration rate values indicate that the filters were not operational therefore rendering the data useless for the modelling of filter effluent variables. As such, the complete data set had a lower mean value and greater variance for filtration rate than that of either modelling data set, resulting in a statistically significant difference. By removing the patterns containing zero values for filtration rate, the modelling data sets are actually more representative of operating conditions than the complete data set.

In order to extract the data patterns into the required training, testing, and production data sets, the data patterns were sorted according to the value of the model output variable. The patterns were then extracted in a 3:1:1 ratio (training:testing:production), the most effective ratio as determined through past modelling experience. Preliminary modelling revealed that a three-layer multi-layer perceptron architecture produced the best results for both the turbidity and particle count models. Other architectures were therefore eliminated from further consideration, a decision supported by previous modelling experience (Stanley *et al.* 2000). For the turbidity model, the best results were obtained when 13 hidden-layer neurons were employed, while for the particle count model, 8 hidden-layer neurons were found to be optimal. When applied to the production data set,

the turbidity model demonstrated excellent predictive capacity, predicting the 50th percentile value of filter effluent turbidity with a mean absolute error of 0.004 NTU and a coefficient of multiple determination (R^2) of 0.85 (Table 3.11). Similarly, the particle count model had an R^2 value of 0.63 and predicted the 50th percentile of filter effluent particle counts with a mean absolute error of 21.9 counts/mL. As was previously discussed, the stability of the models must be ascertained by redistributing the data in the modelling sets and retraining the models using identical modelling variables. This cross-validation technique ensures that model performance is not a function of the manner in which the data sets were extracted. As can be seen in Table 3.12, the cross-validation results for the particle count model shows slight deterioration when compared to those of the original model. For the turbidity model, the results improved slightly during cross-validation. In both cases, the deviation between original model results and cross-validation model results is not great enough to warrant invalidation of either model. Statistical measures, such as a t-test for the difference in two means, can be used to detect a statistically significant difference during model stability analyses.

The modelling results for the turbidity and particle counts models are presented graphically in Figures 3.3 and 3.4 respectively. Due to the limited size of the data sets used in model development, modelling results for each of the training, testing, and production data sets are presented, although the production set results are the most important in determining a model's predictive capacity. Both figures demonstrate that the model provides excellent predictions on the vast majority of the data points. In both cases however, the models have difficulty predicting the largest peaks in the production

data set. These peaks, centered around pattern #69 in each of the data sets, are indicative of sub-optimal treatment process performance as filter effluent variables are at values that exceed those observed in the majority of data patterns. While the models clearly recognize that a peak exists, the absolute error of prediction is greater here than for other data points. This type of prediction error is generally caused by a lack of similar data patterns in the model training set. With regards to the particle count model, for example, there are only 5 patterns that contain a 50^{th} percentile filter effluent particle count value that exceeds 150 counts/mL (Figure 3.4). As only three of these patterns are in the training set, the model has only limited information to draw on when making predictions under sub-optimal operating conditions. The prediction errors of such cases can be improved by increasing the size of the training set to include a greater number of patterns collected during sub-optimal operating conditions. As previously discussed, all of the useable data were employed in model development; increasing the size of data sets is not possible unless further data are collected.

Unfortunately, the models developed for the ODP facility cannot be evaluated online. As a research facility, the operating conditions at the plant are subject to its research schedule. This schedule did not permit the operation of the facility under conditions that mimicked the data used in the models presented herein. In spite of this limitation, it is still possible to ensure that the models are making sensible predictions when applied to process data. As will be discussed in the model applications section, the models can be used as virtual full-scale laboratories to generate process performance graphs. If these

graphs are in agreement with accepted process knowledge, then the validity of the models is further strengthened.

3.4.2.2. F.E. Weymouth Filtration Plant.

Models were developed for both the mean combined filter effluent turbidity and grand mean filter effluent particle counts. With respect to the former, the mean value is calculated on the basis of laboratory grab samples taken over a 24-hour period. The particle count model output is the mean of the 50th percentile value of effluent particle counts for 5 of the facility's 48 filters, those that have online particle count instrumentation. Since the 50th percentile is also known as the median value, the particle counts model output is the mean of median effluent particle counts in the filters; the term grand mean is used to simplify model discussion.

As with the ODP Plant models, inputs for the F.E. Weymouth facility were selected on the basis of data availability and reliability, recent literature, and the results of preliminary modelling, and are listed in Table 3.13. As can be seen in Table 3.8, there are fewer observations for dissolved oxygen concentration than for other turbidity model input variables. After removing patterns lacking dissolved oxygen concentration values as well as those containing other blank entries, as well as erroneous and ambiguous values, the resulting turbidity model data set consists of 712 data patterns spanning three years of plant operations. As was previously discussed, particle count data were only collected during an independent internal study in 1999. As a result the data set is smaller,

consisting of 277 acceptable data patterns. In order to ensure that each of the data sets is representative of the full spectrum of water quality and operational characteristics, ANOVA analyses were performed between each modelling data set and the complete data set. For the turbidity model, there is no statistically significant difference between the modelling data and the complete data set, at the 0.05 level of significance for any variables other than alum dose. Of the 384 data patterns removed when preparing the modelling data set, 378 were removed due to the lack of dissolved oxygen concentration data. The mean alum dose for these removed patterns was 4.39 mg/L, while those for the complete data set and modelling data set were 4.24 mg/L and 4.16 mg/L, respectively. The variances for each of the data sets, presented in the same order, are 0.26 $(mg/L)^2$, 0.30 $(mg/L)^2$, and 0.30 $(mg/L)^2$, respectively. There is no apparent reason for the statistically significant difference; removed patterns are distributed through all seasons and throughout the spectrum of raw water quality conditions. Furthermore, there is no objective way of rectifying the differences in means. As such, the difference is simply noted without further attempts to explain or rectify it.

In stark contrast, there is a statistically significant difference, as determined through an ANOVA analysis at the 0.05 level of significance, between the particle count modelling data set and the complete data set for all input variables except plant influent particle counts, raw water temperature, and effluent particle counts. Since all of the modelling data were collected in 1999, while the complete data set spans from 1997 to 1999, these results suggest that both water quality and operational conditions were significantly different in 1999 than in the two preceding years. Indeed, the modelling data

demonstrated higher mean values of SPW %, plant flow, influent pH, alum dose, polymer dose, as well as lower values of influent turbidity and filtration rate. As only 1999 particle count data are available for modelling, these significant differences in variable means cannot be objectively reduced. Should additional particle count data become available in future years, a more representative data set can be obtained.

As with the ODP Plant models, the data patterns for each of the F.E. Weymouth Filtration Plant models were extracted into training, testing, and production data sets in a 3:1:1 ratio. Preliminary modelling results suggested that the three-layer multi-layer perceptron architecture was once again the most effective architecture for model development, and other architectures were not further pursued. With respect to the number of hidden layer neurons employed in each model, 9 hidden-layer neurons for the turbidity model and 10 for the particle count model were found to be optimal. When applied to the production data set, the turbidity model predicted the mean combined filter effluent turbidity with a mean absolute error of 0.005 NTU and an R^2 value of only 0.23 (Table 3.14). When the data were redistributed and the model was re-trained using identical variables, the modelling results improved slightly (Table 3.15). This cross-validation ensures that model results are not a function of the manner in which the data patterns were extracted into the training, testing, and production sets. The production set results for the turbidity model are presented graphically in Figure 3.5. The model predictions are consistently offset from the actual values. This offset can be explained by the fact that while the model makes output predictions on a continuous scale, the output variable consists of small set of discrete values, as will be discussed.

These results, which are mediocre at best, can best be explained by revisiting the modelling data set. The model output variable, combined filter effluent turbidity, was calculated on the basis of multiple grab samples over a 24 hour period. Although the variable is reported to three decimal places in the data obtained from MWD, the accuracy of the turbidity meters used in sample analysis (±0.01 NTU) dictates that the variable should only be reported to two decimal places. When the combined filter effluent data are truncated to two decimal places, there are only 7 discrete values for the variable, ranging from 0.05 NTU to 0.11 NTU. While it is plausible that there is a real difference between these values caused by water quality and operational variations, it is equally likely that the differences are an artifact resulting from instrumental and averaging errors. In either case, the ANN technique is not well suited for mapping cause effect relationships between countless possible combinations of model inputs and only seven possible output values where the output variable is non-categorical.

The results for the particle count model are far superior to those of the turbidity model. When applied to the production data set, the model predicted the mean filter effluent particle count with a mean absolute error of 7.9 counts/mL and an R^2 of 0.73 (Table 3.14). The cross-validation results indicate that there is little deterioration in model results when the model was re-trained on the redistributed data (Table 3.15). As such, the model's ability to generalize regardless of the manner in which data are extracted is validated. As can be seen in Figure 3.6, the final model shows excellent predictive capacity on previously unseen data. The model recognized the cases where the mean

filter effluent particle counts are greater than normal, indicating sub-optimal process performance, although there is some error in predicting the exact value. Since the goal of process modeling is to develop tools to reduce the occurrence of process upsets, it is generally far better for the model to demonstrate better predictive capacity on the majority of the data than to focus on the few cases where process upsets were evident.

As was mentioned previously, the F.E. Weymouth Filtration Plant SCADA system does not capture values for all water quality variables online in real-time. Plant operations are dependent upon a rigorous water quality sampling and analysis schedule. As such, it is not possible to validate the models online in real-time. The ability of the particle count model to generate useful information and correct predictions will, however, be presented in the applications section. The agreement between model predictions and accepted process trends when the model is used in virtual laboratory and scenario analysis applications can improve user confidence in model predictions and lend validity to model results.

3.5. MODEL APPLICATIONS

This section presents an overview of the utility of ANN models in offline process control applications and in the analysis of small data sets. More specifically virtual laboratory and scenario analysis applications, along with the use of ANN models for operator training, will be discussed. In combination, these applications lend credibility to the ANN

modelling technique and its further implementation in the drinking water treatment industry.

3.5.1. Virtual Laboratory and Scenario Analysis

Trained ANN models can be used as a powerful tool for the analysis of complex treatment scenarios in order to improve future operations. This same tool can be used to train operators by allowing them to conduct offline exercises that emphasize efficient control of difficult-to-manage scenarios. Alternatively, the trained models can be used as virtual full-scale laboratories to gain insight into the important factors that affect a particular process. In combination, these tools allow for more educated and efficient control decisions, particularly during conditions where maintaining optimal process control is difficult.

3.5.1.1. Methodology and Results

3.5.1.1.1. Virtual Laboratory Applications

Virtual laboratory applications harness the ANN modelling technique's ability to capture the cause-effect relationships that exist between a multitude of inputs and a single output variable. For the purposes of the examples described herein, the trained ANN models were interfaced with Microsoft Excel spreadsheets as described in detail in Chapter 6.

Once the interface is established, model predictions can be generated for any reasonable user-defined data pattern, as manually entered in the spreadsheet.

The F.E. Weymouth Filtration Plant particle count model was interfaced to an Excel spreadsheet to form a virtual laboratory. The application was used to determine the effects of several water quality variables on typical summer operations. In order to ensure a successful application, representative values for each of the model inputs must be selected. For this particular case, mean values for the month of August 1999 were used as model variables and are listed in Table 3.16. Many of the variables have a very narrow range over the selected time frame, as emphasized by the 10^{th} and 90^{th} percentile values. In evaluating the effects of various variables within these ranges, meaningful virtual laboratory results can be obtained.

The present virtual laboratory analysis involves the determination of the effects of plant influent turbidity and plant influent particle counts on the value of filter effluent particle counts. As can be seen in Figure 3.7, as the value of plant influent particle counts increases over the specified range, the value of filter effluent particle counts also increases. This effect is rather insignificant when compared to the effect of plant influent turbidity. For typical summer water, process performance actually improves as turbidity increases over the specified range. All August 1999 data patterns had the same chemical dosing conditions; 4.5 mg/L of alum and 1.5 mg/L of polymer. When the influent turbidity is lower, there is insufficient substrate to form effective flocs and the coagulant is carried over to the filters. Of all the raw water quality variables, pH appears to have

the greatest impact on coagulation performance, as measured by filter effluent particle counts (Figure 3.8). The predominant mechanism for coagulation at MWD facilities is charge neutralization, as evidenced by low alum doses. As no chemicals are added for coagulation pH control, relatively small increases in influent pH can upset the pH of the mixed coagulation solution, resulting in comparatively large variations in process performance.

The virtual laboratory technique can also be used to determine the effects of chemical doses and other process conditions on process performance, as an alternative to bench-scale jar-test type analyses. For typical summer water, as defined in Table 3.16, the effect of alum and polymer doses is presented graphically in Figure 3.9. Both alum and polymer are effective at reducing filter effluent particle counts over the specified ranges. The effect of polymer dose is most prominent when lower doses of alum are applied. As the alum dose is increased and filter effluent particle counts are minimized, increasing polymer dose has little effect. The most efficient alum and polymer dose combination for a given water type can be determined using a similar plot in combination with chemical cost information. In controlling the process, the operator also has the option of altering the values of plant flow and filtration rate.

As can be seen in Figure 3.10, plant flow has a tremendous impact on filtration performance, while the effect of filtration rate over the specified range is minimal. Furthermore, it appears as though the least optimal plant flow is around 330 MGD, with lower filter effluent particle counts observed on either side of this value. According to

operations staff at the F.E. Weymouth facility, at certain flows water cascades through parts of the pipeline thereby increasing dissolved gas concentrations. Higher dissolved gas concentrations can cause filter air binding and negatively impact filter performance. When the plant flow exceeds a threshold value, the upper feeder line runs full and introduces less dissolved gas. The numerical value of this flow threshold, as well as the flow which corresponds to decreased particle removal, is dependent upon the number of treatment trains and filter basins that are in operation.

3.5.1.1.2. Scenario Analysis Applications.

In scenario analysis applications, trained ANN models are used to study a particular operational event and determine how it is affected by different process conditions. These applications are particularly useful in determining the cause of process upsets, as well as establishing operational heuristics for future similar events. As with the virtual laboratory applications, the ANN model is interfaced with a spreadsheet, however, the model inputs are taken from the event being studied, rather than being mean or other calculated values. For demonstration purposes, a sub-optimal treatment event at the ODP Plant has been identified. On August 26th, 1997 the values of the input variables presented in Table 3.17 resulted in a filter effluent particle count value of 56.3 counts/mL. Using the ODP Plant particle count model, in combination with the input data from the event, it is possible to determine if different operational decisions would have led to better treatment performance. As is demonstrated in Figure 3.11, the 4.3 mg/L ferric chloride dose applied was far beyond the optimal value for the particular raw

water observed during the event. By increasing the polymer dose from 2.9 mg/L to 3.1 mg/L while scaling back the ferric chloride dose to 2.5 mg/L, a filter effluent particle count value of approximately 7.6 counts/mL was predicted. The predicted solution may allow for a more efficient use of chemicals while improving finished water quality.

3.5.1.1.3. Operator Training.

The scenario analysis technique is particularly useful in continuing operator training or in the training of new operators. On September 10th 1999, the combination of water quality and operational variable values resulted in an excessively high mean 50th percentile filter effluent particle count value of 92.3 counts/mL at the F.E. Weymouth Filtration Plant. As a training exercise, operators are given access to a custom scenario analysis application where the values of all the input variables except alum dose, polymer dose, and plant flow have been set to the values recorded on September 10th. By manipulating the chemical dosing levels as well as the plant flow, operators can determine the effects of control decisions on filtered water quality. This particular event is difficult to optimize, as even with high doses of alum and polymer, 5.5 mg/L and 1.75 mg/L respectively, the filter effluent particle counts remain relatively high, approximately 26.5 counts/mL. Through this simple exercise, the operator learns that the only way to effectively lower the filter effluent particle counts is to alter the plant flow. As is demonstrated in Figure 3.12, both lowering and increasing the plant flow for this particular combination of raw water quality and dosing conditions can reduce the filter

effluent particle counts. The best option can be selected by balancing filtration performance with customer water demand.

3.6. CONCLUSIONS

The results presented herein suggest that the ANN modelling technique is applicable to a wide variety of water treatment process data. In particular, successful models were developed using relatively small data sets. It is important to note, however, that not all data are amenable to process modelling using the ANN technique. As was demonstrated through the F.E. Weymouth Filtration Plant turbidity model results, sub-optimal model performance can be expected when the model output variable is a continuous random variable that has a limited range of values in the modelling data set.

The applications developed using models of MWD facilities have demonstrated an excellent ability to provide useful operational information so that operators can make more educated control decisions. The virtual laboratory applications serve to highlight important water quality and process variables, as well as the effects on the process output. Scenario analysis applications allow operators to analyze operational events to improve future process performance. When used alone or in combination with the applications discussed in Chapters 5 and 6, these tools can substantially improve operational efficiency.

3.7. REFERENCES

Stanley, S.J., Baxter, C.W., Zhang, Q., and Shariff, R. 2000. *Process Modelling and Control of Enhanced Coagulation.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO: 167 p.

Table 3.1 Water quality variables measured during the filtration study at the ODP Plant

Variable	Location
Source water blend (% SPW)	Plant Influent
Turbidity (NTU)	Plant Influent, filter influent, filter effluent[1]
Particle counts	Filter effluent[1]
pH	Plant Influent, filter influent, filter effluent
Dissolved oxygen (mg/L)	Plant Influent
UV_{254} absorbance(cm^{-1})	Plant Influent
Temperature (°C)	Plant Influent
Alkalinity (mg/L)	Plant Influent
Bromide (mg/L)	Plant Influent

[1]Collected using online instruments

Table 3.2 Operational variables measured during the filtration study at the ODP Plant

Variable	Type of Measurement
Plant flow (MGD)	Operational set-point
$FeCl_3$ dose (mg/L)	Operational set-point
Polymer dose (mg/L)	Operational set-point
Filter-aid dose (mg/L)	Operational set-point
Total ozone dose (mg/L)	Operational set-point
Filtration rate (gpm/ft^2)	Operational set-point
Length of filter run (h)	Direct measurement
Filter headloss (ft)	Calculated value

Table 3.3 Water quality variables measured at the F.E. Weymouth Filtration Plant

Variable	Sample Location(s)
Free chlorine (mg/L)	Plant influent, flocculation basin influent, clarifier effluent, filter effluent, reservoir influent, plant effluent
Total ammonia nitrogen (mg/L)	Plant influent, flocculation basin influent, filter effluent, reservoir influent, plant effluent
Total chlorine (mg/L)	Plant influent, flocculation basin influent, filter effluent, reservoir influent, plant effluent
Turbidity (NTU)	Plant influent, flocculation basin influent, combined filter effluent, plant effluent
pH	Plant influent, flocculation basin influent, clarifier effluent, filter effluent, reservoir influent, plant effluent
Dissolved oxygen (mg/L)	Plant influent, plant effluent
Odor (TON)	Plant influent, plant effluent
Temperature (°C)	Plant influent
Total hardness (mg/L as $CaCO_3$)	Plant influent
Source water blend (% SPW)	Plant influent
Free ammonia nitrogen (mg/L)	Flocculation basin influent, filter effluent, plant effluent
Monochloramines (mg/L)	Flocculation basin influent, filter effluent, plant effluent
Dichloramines (mg/L)	Flocculation basin influent, filter effluent, plant effluent

Table 3.4 Operational variables measured at the F.E. Weymouth Filtration Plant

Variable	Type of Measurement
Plant flow (MGD)	Operational set-point
Filtration rate (gpm/ft^2)	24 hour average value
Number of filters in service	24 hour average value
Length of filter run (h)	24 hour average value
Total headloss (ft)	24 hour average value
Ammonia dose (mg/L)	Operational set-point
Chlorine dose at plant influent (mg/L)	Operational set-point
Chlorine dose at filter influent (mg/L)	Operational set-point
Chlorine dose at plant effluent (mg/L)	Operational set-point
Filter-aid dose (mg/L)	Operational set-point
Polymer dose (mg/L)	Operational set-point
Alum dose (mg/L)	Operational set-point
Caustic soda dose (mg/L)	Operational set-point

Table 3.5 Variables measured by online instruments at the F.E. Weymouth Filtration Plant

Variable	Location
Chlorine residual (mg/L)	Plant influent, middle of flocculation basin, filter effluent, reservoir influent, plant effluent
pH	Plant influent, reservoir influent, plant effluent
Dissolved oxygen (mg/L)	Plant influent, plant effluent
Turbidity (NTU)	Plant influent, filter effluent, reservoir influent, plant effluent
Temperature	Plant influent, plant effluent
Ammonia (mg/L)	Reservoir influent

Table 3.6 ODP Plant, data analysis of raw water quality variables

Variable	# Obs.	Mean	Std. Dev.	Min	Max	Percentile 5th	Percentile 50th	Percentile 95th
Temperature (°C)	107	19.50	4.10	12.22	25.11	12.52	20.56	24.57
Turbidity (NTU)	99	3.53	1.91	1.08	9.60	1.30	3.00	7.30
pH	98	8.12	0.22	7.65	9.07	7.84	8.05	8.39
Dissolved oxygen (mg/L)	93	8.00	1.74	4.28	15.00	5.67	7.74	10.60
UV_{254} absorbance (cm^{-1})	16	0.07	0.02	0.04	0.10	0.04	0.08	0.09
Alkalinity (mg/L)	16	76.5	22.1	62.0	135.0	62.0	68.0	133.5
Bromide (mg/L)	16	0.13	0.02	0.09	0.16	0.10	0.14	0.16

Table 3.7 ODP Plant, data analysis of operational and filter effluent variables

Variable	Mean	Std. Dev.	Min	Max	Percentile 5th	50th	95th
Flow (MGD)*	4.80	0.38	3.73	5.64	4.50	4.61	5.49
$FeCl_3$ dose (mg/L)	4.17	1.29	1.67	7.10	2.14	4.36	5.91
Polymer dose (mg/L)	2.67	0.72	1.60	3.92	1.65	2.88	3.83
Filter-aid dose (mg/L)	0.01	0.00	0.00	0.01	0.01	0.01	0.01
Total O_3 dose (mg/L)	1.13	0.34	0.52	1.67	0.57	1.24	1.57
Filtration rate (gpm/ft^2)**	3.83	1.17	0.00	4.90	0.00	3.99	4.90
Turbidity (50^{th} percentile, NTU)	0.05	0.01	0.03	0.10	0.03	0.04	0.07
Turbidity (90^{th} percentile, NTU)	0.06	0.03	0.03	0.28	0.04	0.05	0.08
Particle counts (50^{th} percentile, mL^{-1})	46.74	43.24	4.50	202.70	5.90	18.5	146.09
Particle counts (90^{th} percentile, mL^{-1})	168.02	239.78	22.20	2392.4	27.28	69.1	366.64

* note: 1 MGD = 3.754 ML/d

** note: 1 gpm/ft^2 = 2.445 $(m^3/h)/m^2$

Table 3.8 F.E. Weymouth Filtration Plant, data analysis of raw water quality variables

Variable	# Obs.	Mean	Std. Dev.	Min	Max	Percentile 5th	50th	95th
Source blend (% SPW)	1093	15.45	14.07	0.00	100.00	0.00	20.83	31.4
Turbidity (NTU)	1087	1.82	0.70	0.51	7.23	0.93	1.68	3.96
pH	1087	8.25	0.10	7.67	8.49	8.03	8.27	8.36
Dissolved oxygen (mg/L)	719	8.0	1.1	5.8	12.0	6.2	8.0	9.8
Temperature (°C)	1084	18.0	4.4	10.4	25.4	12.6	17.2	25.0
Particle counts (10^{th} pct., mL^{-1})	281	3256	1067	661	9595	1875	3197	7591
Particle counts (50^{th} pct., mL^{-1})	281	3672	1209	843	10873	2188	3565	8700
Particle counts (90^{th} pct., mL^{-1})	281	5227	6063	946	56054	2525	4056	39710

Table 3.9 F.E Weymouth Filtration Plant, data analysis of operational and filter effluent variables

Variable	Mean	Std. Dev.	Min	Max	Percentile 5th	50th	95th
Flow (MGD)*	252	51	93	380	180	250	335
Alum dose (mg/L)	4.24	0.55	3.50	6.50	3.50	4.25	5.50
Polymer dose (mg/L)	1.52	0.18	1.30	3.83	1.30	1.50	1.75
Free chlorine dose (Plant influent, mg/L)	2.66	0.37	0.52	3.00	2.00	2.75	3.00
Free chlorine dose (Filter influent, mg/L)	1.48	0.46	0.39	2.91	0.80	1.41	2.28
Filtration rate (gpm/ft^2)**	2.93	0.17	2.01	3.97	2.68	2.92	3.22
Turbidity (NTU)	0.07	0.01	0.05	0.11	0.06	0.07	0.08
Particle counts (10^{th} percentile, mL^{-1})	8.0	15.8	0.3	329.6	0.7	2.9	31.4
Particle counts (50^{th} percentile, mL^{-1})	13.8	25.7	0.3	396.2	1.2	4.7	59.0
Particle counts (90^{th} percentile, mL^{-1})	24.0	42.8	0.3	485.8	2.1	8.4	105.7

* note: 1 MGD = 3.754 ML/d

** note: 1 gpm/ft^2 = 2.445 $(m^3/h)/m^2$

Table 3.10 ODP Plant, model input variables

50^{th} Percentile Filter Effluent Turbidity Model	50^{th} Percentile Filter Effluent Particle Counts Model
Plant influent temperature (°C)	Plant influent turbidity (NTU)
Plant influent turbidity (NTU)	Plant influent pH
Plant influent pH	$FeCl_3$ dose (mg/L)
$FeCl_3$ dose (mg/L)	Polymer dose (mg/L)
Filter-aid dose (mg/L)	Filter-aid dose (mg/L)
Filtration rate (gpm/ft^2)	Total ozone dose (mg/L)
Filter influent turbidity (NTU)	Filtration rate (gpm/ft^2)

Table 3.11 ODP Plant, modelling results

Model	R^2	MAE
50^{th} percentile filter effluent turbidity	0.85	0.004 NTU
50^{th} percentile filter effluent particle counts	0.63	21.9 counts/mL

Table 3.12 ODP Plant, model cross-validation results

Model	R^2	MAE
50th percentile filter effluent turbidity	0.90	0.003 NTU
50th percentile filter effluent particle counts	0.56	22.6 counts/mL

Table 3.13 F.E. Weymouth Filtration Plant, model input variables

Mean Combined Filter Effluent Turbidity Model	Grand Mean Filter Effluent Particle Counts Model
State Project Water blend (%SPW)	State Project Water blend (%SPW)
Plant influent dissolved oxygen (mg/L)	50th percentile plant influent particle counts (counts/mL)
Plant influent temperature (°C)	Plant influent temperature (°C)
Plant influent turbidity (NTU)	Plant influent turbidity (NTU)
Plant influent pH	Plant influent pH
Alum dose (mg/L)	Alum dose (mg/L)
Polymer dose (mg/L)	Polymer dose (mg/L)
Filtration rate (gpm/ft^2)	Filtration rate (gpm/ft^2)
Plant flow (MGD)	Plant flow (MGD)

Table 3.14 F.E. Weymouth Filtration Plant, modelling results

Model	R^2	MAE
Combined filter effluent turbidity	0.23	0.005 NTU
Grand mean filter effluent particle counts	0.73	7.9 counts/mL

Table 3.15 F.E. Weymouth Filtration Plant, model cross-validation results

Model	R^2	MAE
Combined filter effluent turbidity	0.25	0.005 NTU
Grand mean filter effluent particle counts	0.59	8.1 counts/mL

Table 3.16 F.E. Weymouth Filtration Plant, data used in particle count virtual laboratory development

Variable	Mean	Percentile 10th	Percentile 90th
State Project Water (%SPW)	25.9	25.3	26.5
Plant influent particle counts (50^{th} percentile, counts/ml)	3418.6	2971.3	3943.2
Plant flow (MGD)*	326.5	315.9	340.0
Plant influent turbidity (NTU)	2.2	2.0	2.4
Plant influent pH	8.2	8.2	8.2
Plant influent temperature (°C)	23.5	22.9	23.9
Alum dose (mg/L)	4.5	4.5	4.5
Polymer dose (mg/L)	1.5	1.5	1.5
Filtration rate (gpm/ft^2)**	2.8	2.7	2.9

* note: 1 MGD = 3.754 ML/d

** note: 1 gpm/ft^2 = 2.445 $(m^3/h)/m^2$

Table 3.17 ODP Plant, scenario analysis event data

Variable	Value, August 26^{th} 1997
Plant influent turbidity (NTU)	3.8
Plant influent pH	8
$FeCl_3$ dose (mg/L)	4.3
Filter- aid dose (mg/L)	0.01
Polymer dose (mg/L)	2.9
Filtration rate (gpm/ft^2)*	3.99
Total ozone dose (mg/L)	1.3

* note: 1 gpm/ft^2 = 2.445 $(m^3/h)/m^2$

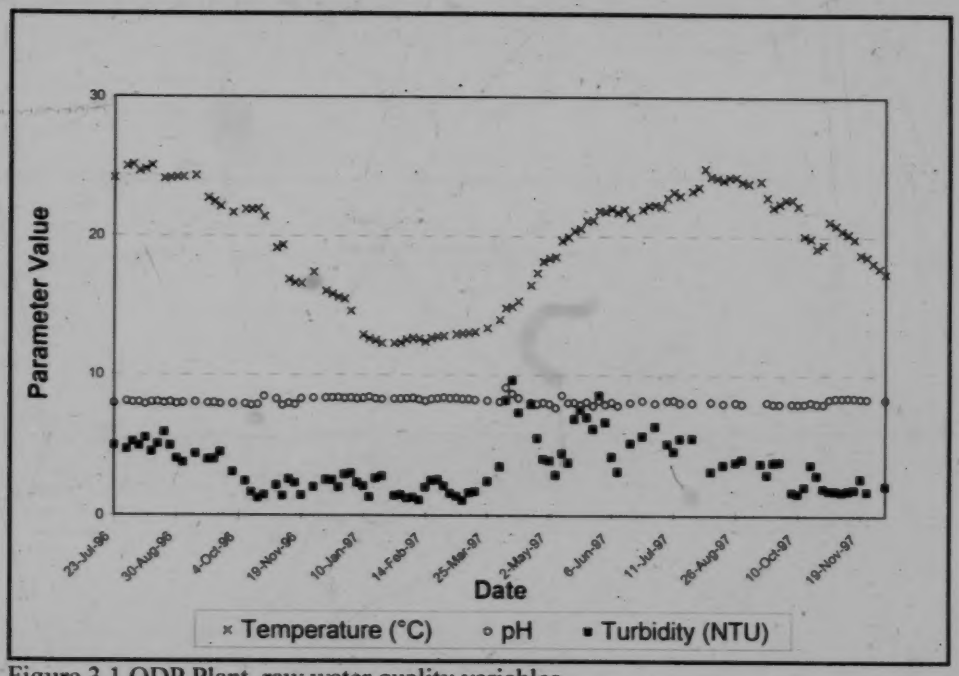


Figure 3.1 ODP Plant, raw water quality variables

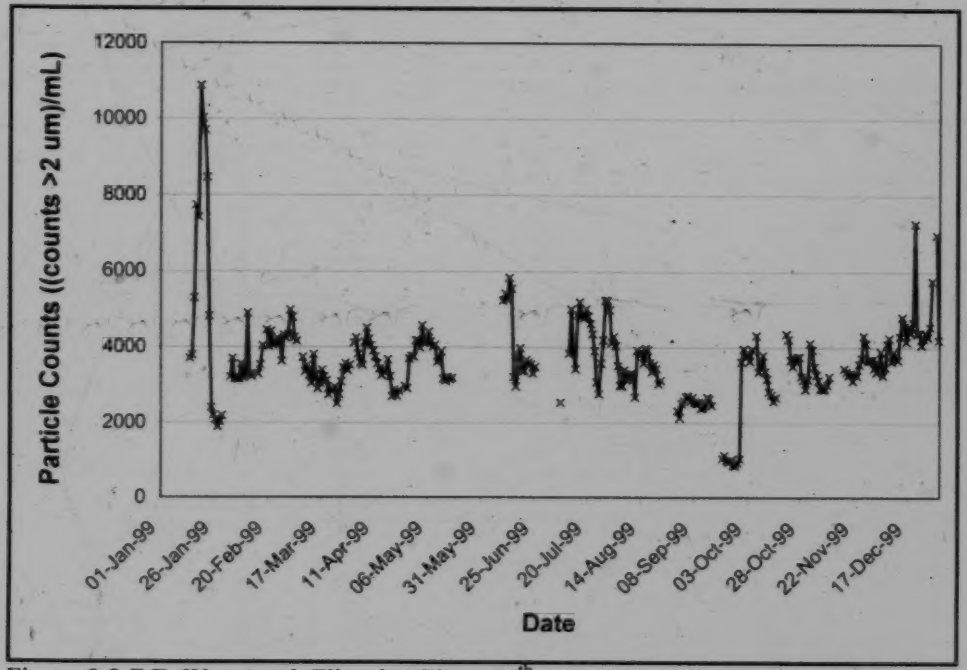


Figure 3.2 F.E. Weymouth Filtration Plant, 50th percentile plant influent particle counts

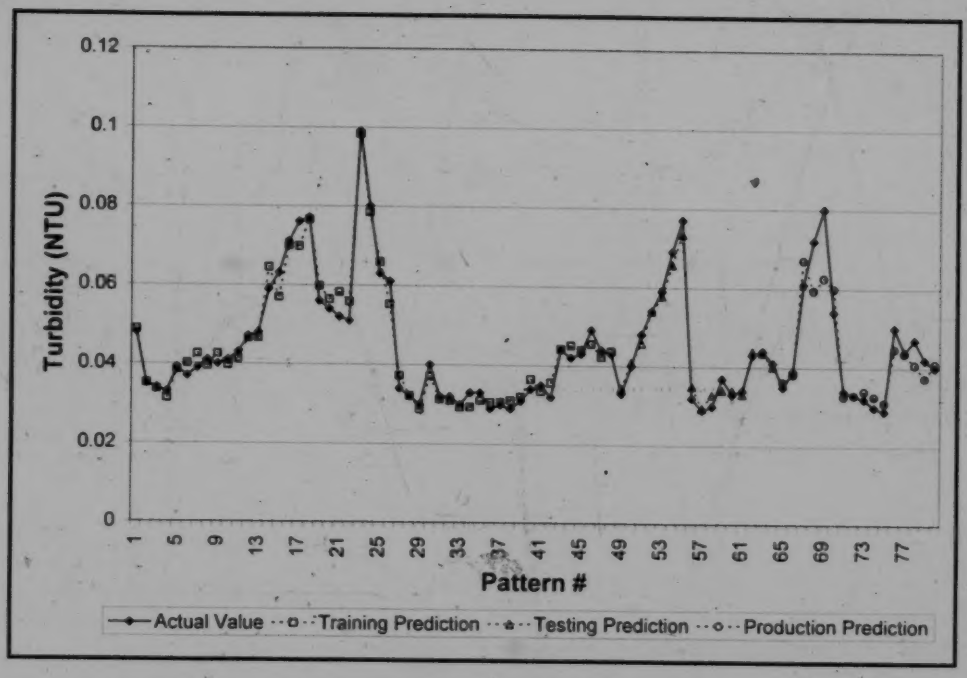


Figure 3.3 ODP Plant, 50th percentile filter effluent turbidity model results

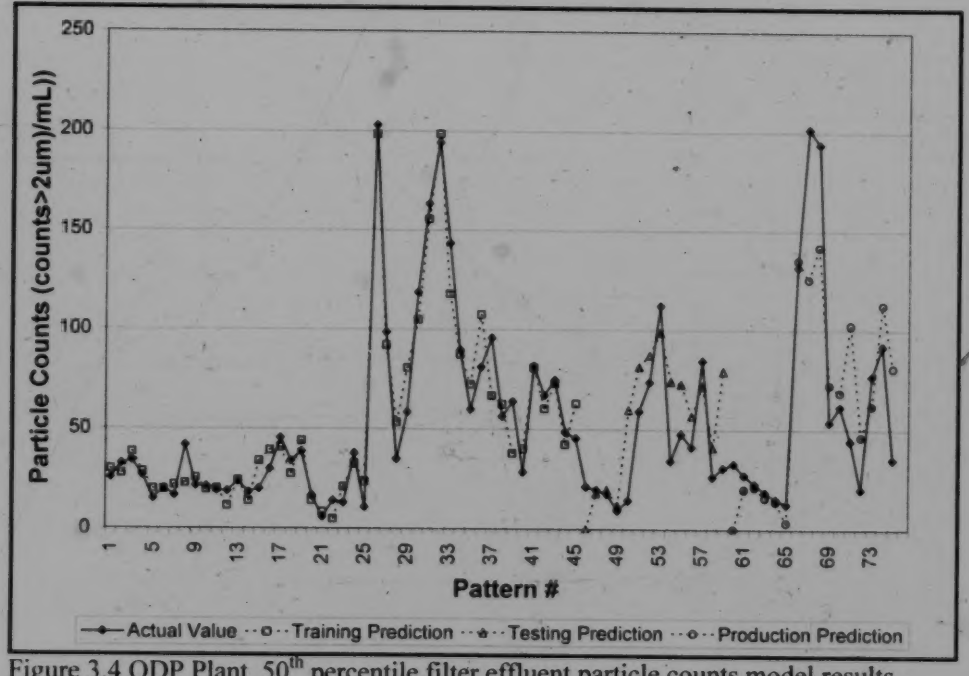


Figure 3.4 ODP Plant, 50th percentile filter effluent particle counts model results

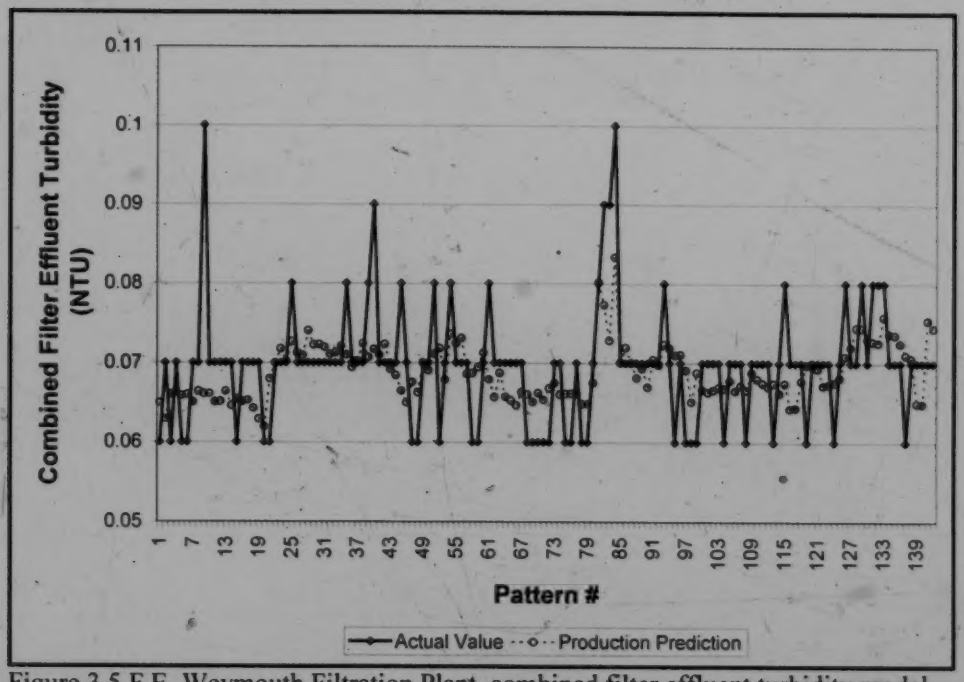


Figure 3.5 F.E. Weymouth Filtration Plant, combined filter effluent turbidity model results

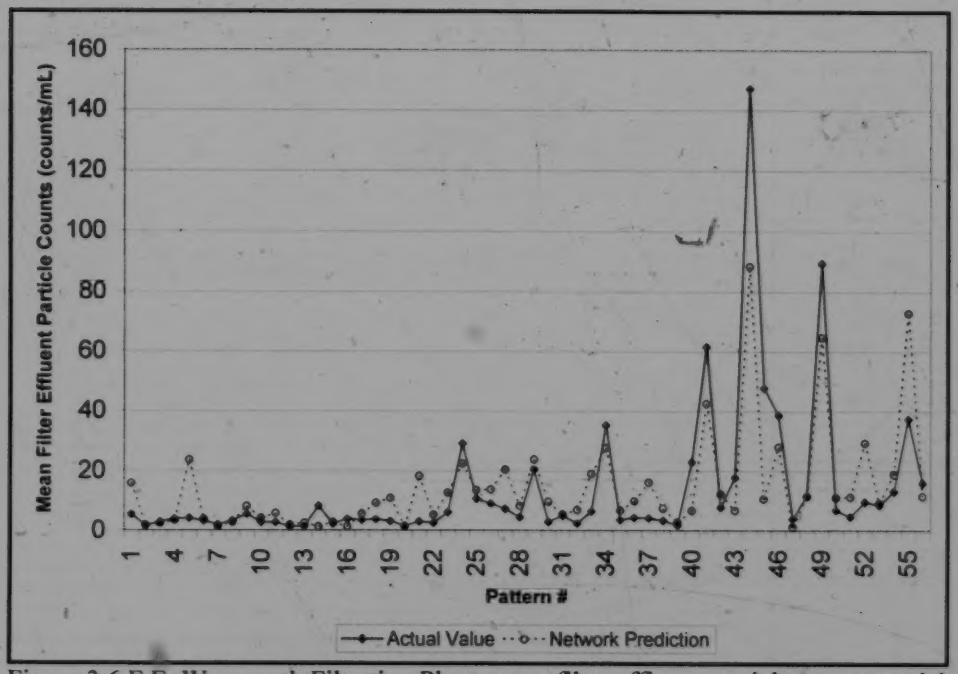


Figure 3.6 F.E. Weymouth Filtration Plant, mean filter effluent particle counts model results

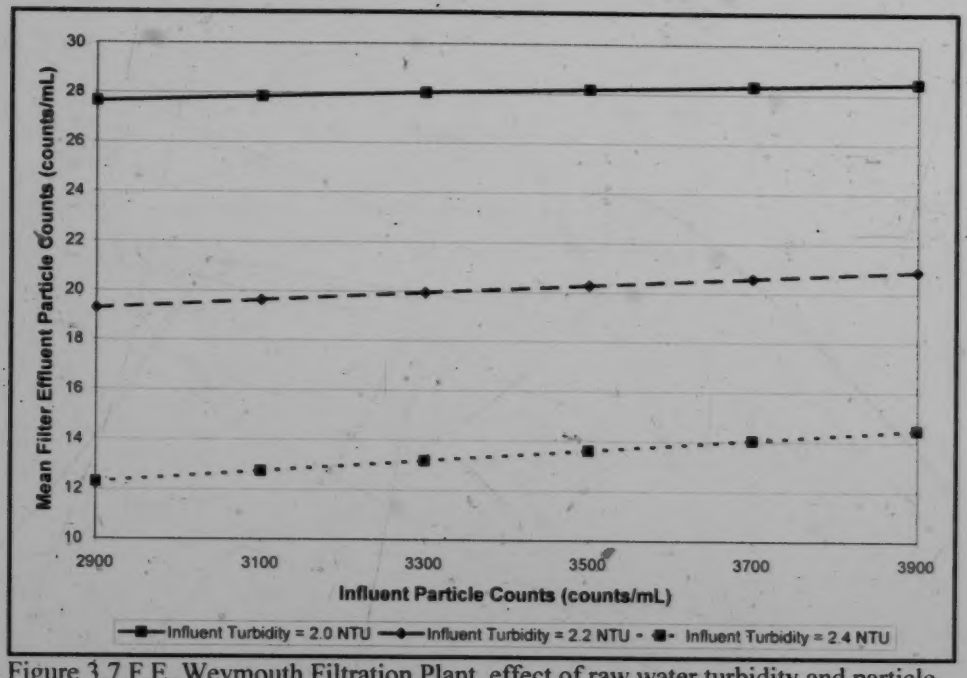


Figure 3.7 F.E. Weymouth Filtration Plant, effect of raw water turbidity and particle counts on filter effluent particle counts for typical summer water

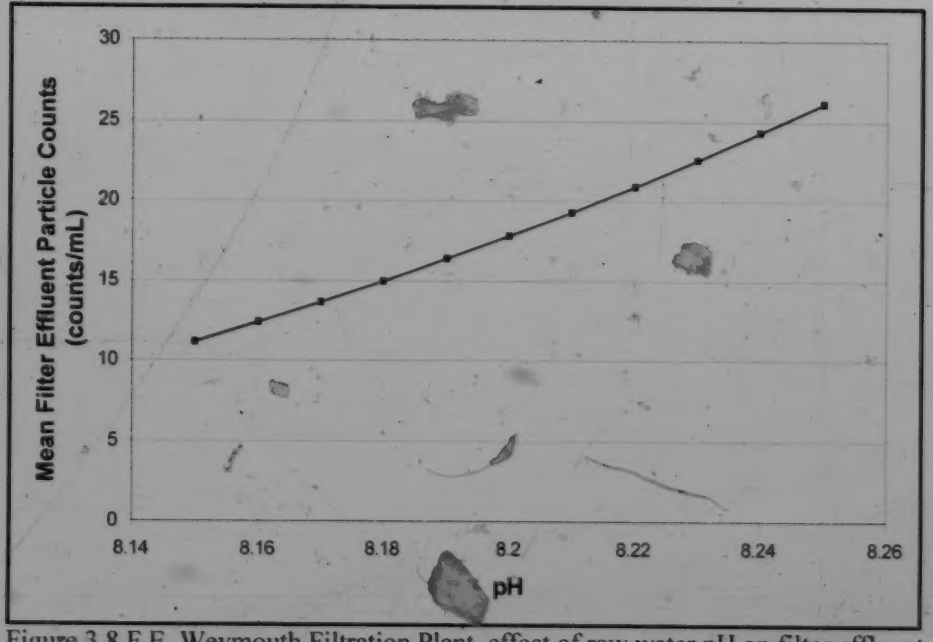


Figure 3.8 F.E. Weymouth Filtration Plant, effect of raw water pH on filter effluent particle counts for typical summer water

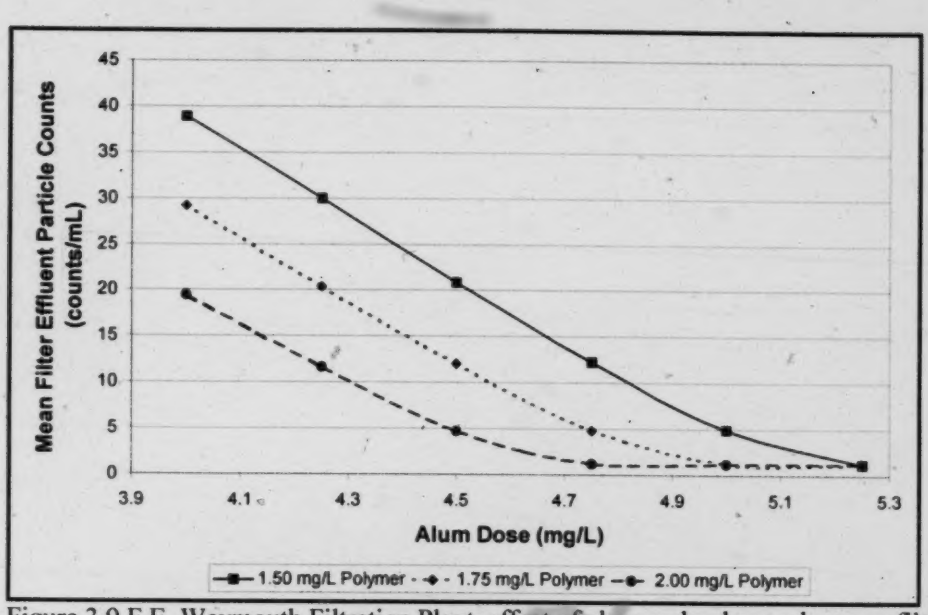


Figure 3.9 F.E. Weymouth Filtration Plant, effect of alum and polymer doses on filter effluent particle counts for typical summer water

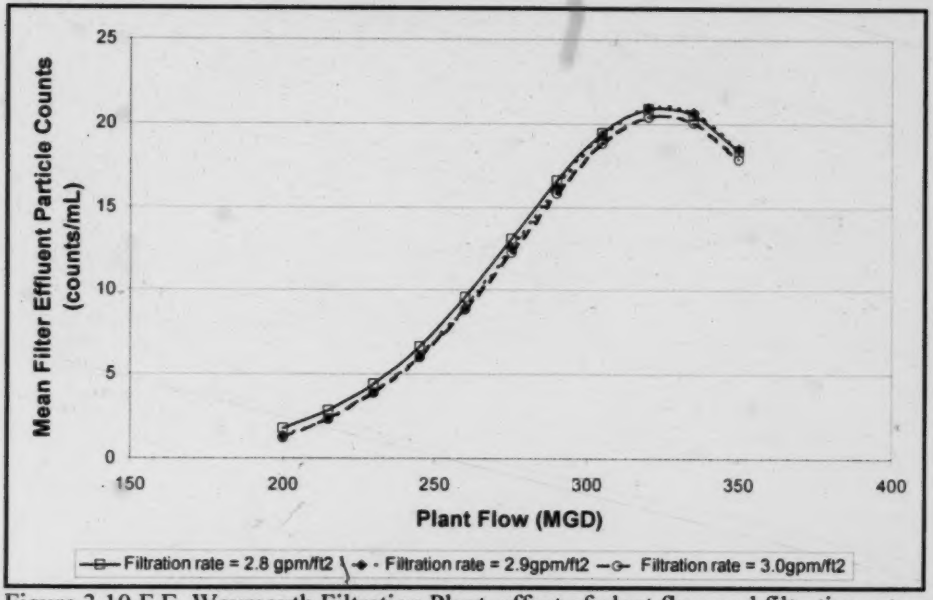


Figure 3.10 F.E. Weymouth Filtration Plant, effect of plant flow and filtration rate on filter effluent particle counts for typical summer water

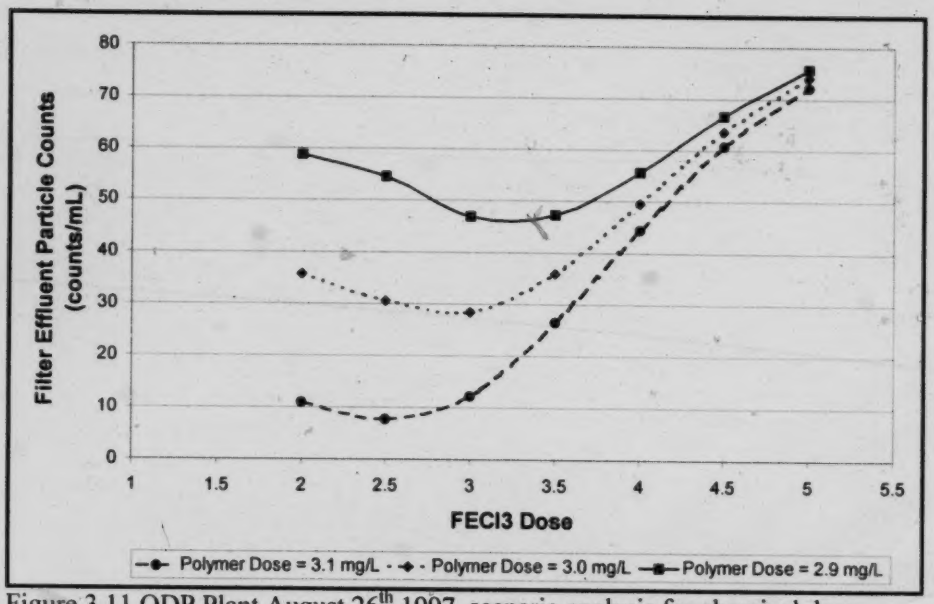


Figure 3.11 ODP Plant August 26th 1997, scenario analysis for chemical dose optimization

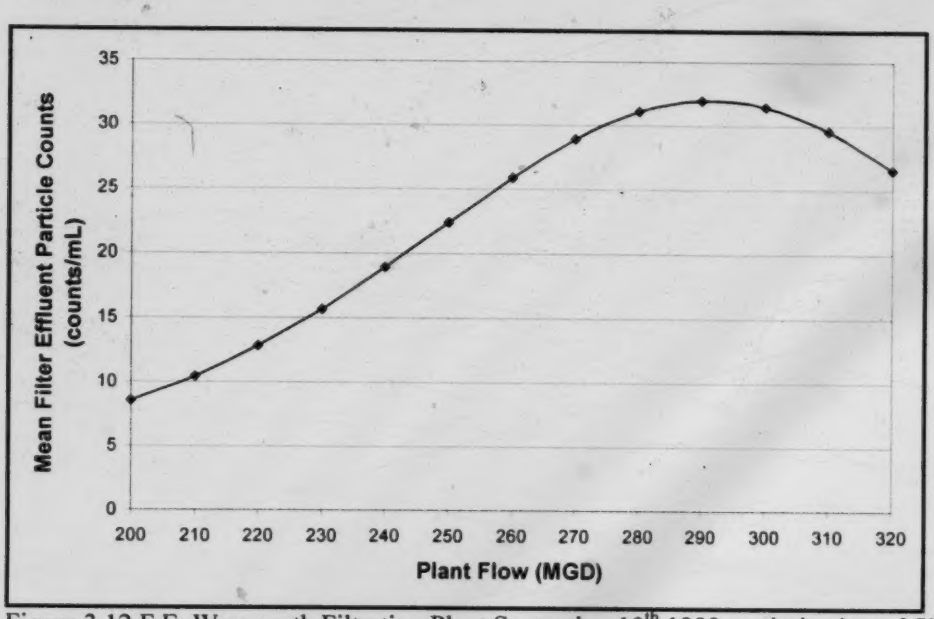


Figure 3.12 F.E. Weymouth Filtration Plant September 10th 1999, optimization of filter performance through plant flow variation

4. EVALUATION OF MODEL BOUNDARIES*

4.1. INTRODUCTION

One of the main concerns surrounding the application of ANN models in process control applications is the determination of model prediction boundaries. ANNs interpolate well within their training domain; accurate predictions can be made as long as the values of each of the input variables falls within the range of values on which the models were developed. Little is known, however, about the ability of ANN models to provide accurate predictions for data outside this training domain. It has been hypothesized that the ANN models will continue to provide reasonably accurate predictions up to a certain critical distance outside the boundaries of the training domain. Beyond this critical distance, ANN models will invariably break down and provide erroneous predictions.

This chapter examines the ability of models to generate accurate predictions beyond their training domain, as well as the impacts of key model training parameters on model prediction boundaries. With respect to the former, the ability of trained models to make accurate predictions when either one model input variable goes out of range, or an entirely new set of raw water quality and operational conditions is presented to the model, is explored. In the case of the latter, the impacts of software-specific training subtleties on model prediction boundaries are discussed. The results of these studies will

* A version of this chapter has been published. Baxter, C.W., Tupas, R-R.T., Zhang, Q., Shariff, R., Stanley, S.J., Coffey, B.M., and Graff, K.G. 2001. *Artificial Intelligence Systems for Water Treatment Plant Optimization.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 141 p.

serve to improve understanding of model extrapolation capabilities, as well as the impacts of new data and training parameters on model performance.

4.2. BACKGROUND INFORMATION

In the last decade, there has been a significant increase in the application of the ANN technique in process modelling and control in the drinking water treatment industry. As automation becomes a part of standard operating procedures in the industry, utilities will look to the ANN technique to serve as a foundation in model-based advanced process control schemes, as discussed in Chapter 6. When integrating models into process control schemes, an understanding of the models' behaviour outside the training domain is essential in order to set effective alarms.

4.2.1. Expanded Data Domain Evaluation

One of the key concerns surrounding the application of ANN models in process control applications is the ability of the models to handle data outside of the training domain. Predictions in such an expanded data domain are crucial when significant future variations in raw water quality are possible. If, for example, a water treatment facility uses a river water source, infrequent events such as 1-in-10-year floods can create raw water quality conditions that were previously unseen. Another situation where out-of-range data might be expected is when ANN models are developed on existing data that are not fully representative of all of the raw water quality and operational conditions that

are expected at the facility. The common practice when developing ANN models where data is limited is to closely monitor model predictions and update the model as deficiencies are detected and new data becomes available. An example of this cycle of model development and model updating is presented in Chapter 6 where models were originally developed on a limited set of data collected during early winter conditions and were updated as raw water quality characteristics changed. Knowledge about model prediction abilities under such situations is crucial if the ANN technology is to see increased application in water treatment process control. Plant operators need to have confidence in model prediction boundaries so that effective alarms can be set in ANN-based control systems.

4.2.2. Evaluation of Scaling Effects

Each ANN modelling software package has its own model development nuances that can result in differences in models' abilities to provide accurate predictions beyond the training domain. ANN modelling requires that all input data variables are scaled from their numeric range to a common range, typically [0,1] or [-1,1] depending on the nature of the data. In most software packages, the minimum and maximum values of each variable are mapped to the minimum and maximum values of the scaled-down range. Scaling raw water temperature to a range of [0,1] would result in the lowest temperature value being scaled to 0 and the highest being scaled to 1, for example. Within these boundaries, all remaining data are scaled, typically using a linear function, although other non-linear functions are possible. Another scaling factor that can affect the model's

performance outside the training domain is whether the scaling range is "open" or closed". In open scaling, when the model is applied to previously unseen data where one or more variables have values that extend beyond the training domain, the scaled values are allowed to fall outside the scaling range. In closed scaling, values that exceed the minimum or maximum values of the training domain are scaled to the minimum or maximum value of the scaling range. Take, for example, a model of clarifier effluent turbidity that has raw water turbidity with a range of 0 NTU to 100 NTU as a model input variable. If the model is scaled linearly to a range of [0,1], a turbidity of 0 would be scaled to 0 and a turbidity of 100 would be scaled to 1. If the model is then applied to a data pattern where the raw water turbidity is 120 NTU, this value would be scaled to either 1 or 1.2 depending on whether closed or open scaling, respectively, was employed. A final factor affecting model predictions outside the training domain that is related to scaling is the ability of some software packages to artificially increase the minimum and maximum values of a data variable by a certain absolute value or percentage so that future values that exceed the data range will be accommodated in the scaled range.

4.3. METHODOLOGY

4.3.1. Expanded Data Domain Evaluation

The data used in the expanded data domain studies were obtained from EPCOR Water Services' Rossdale and E.L. Smith Water Treatment Plants in Edmonton, Alberta, Canada. The data were previously used in historical model development as presented by

Stanley *et al.* (2000). Two separate sets of models were developed. The first set was developed in order to determine the impact of a single model input parameter going out of range, while the second set was used to evaluate the impact of applying models trained during a limited time frame to data outside of the time frame. All models were developed on at least two years of operational data using Statsistica Neural Networks and the modified protocol presented in Chapter 3.

The first set of models consisted of three individual models, one model for the prediction of clarifier effluent turbidity and two for the prediction of clarifier effluent colour. For the turbidity model, the data were obtained from the Rossdale WTP and the model inputs consisted of raw water quality, operational, and time series variables. Model results, as originally presented suggested that raw water turbidity was among the most important model inputs (Stanley *et al.* 2000). The value of this variable, which had a range of 2 NTU to 1481 NTU, was used to classify the data into either a modelling data set or an expanded domain data set. The 75^{th} percentile value of raw water turbidity, equivalent to 31 NTU, was selected as the boundary between the two data sets. All patterns with a raw water turbidity ≤ 31 NTU were used to develop an ANN model, while the other patterns formed the expanded data domain data. The inputs for the colour models also consisted of raw water quality, operational, and time series parameters. Among these inputs, plant influent colour was found to be one of the most important parameters (Stanley *et al.* 2000). Two separate models were developed, using data from the E.L. Smith WTP, where the value of plant influent colour was truncated at the 75^{th} and 95^{th} percentile

values. The ability of the models to generate predictions on out-of-range data was determined by applying each of these models to the remaining data.

The second set of models consisted of two models for the prediction of clarifier effluent colour. The data used in their development, obtained from the E.L. Smith WTP, of the models presented by Stanley *et al.* (2000) were categorized into seven raw water quality classes using Kohonen ANNs available in the NeuroShell2 software package. The classification technique, discussed briefly in Chapter 2, allows for the separation of a set of data into a user-defined number of classes according to the Euclidean distance between the data points in N-dimensional space, where N is the number of inputs used for categorization. Following classification, one model was built using winter raw water quality data only, while another was built using summer raw water quality data. These models were then applied to data from other seasons in order to determine impacts of new water quality and operational conditions on model performance.

4.3.2. Evaluation of Scaling Effects

The scenario analysis techniques discussed in Chapter 3 were applied to both the F.E. Weymouth Filtration Plant and ODP Plant filter effluent particle count models in order to determine the effects of open and closed scaling, as well as variable range manipulation. The NeuroShell2 software package is truly flexible with regards to scaling. Both the open and closed paradigms are supported, and the minimum and maximum values of any

variable or group of variables can be manipulated by the user. In contrast, StatSoft's Statistica Neural Networks supports only the open scaling paradigm.

4.4. RESULTS AND DISCUSSION

4.4.1. Expanded Data Domain Evaluation

As previously discussed, ANN models were developed to determine the impacts of both a single model input parameter exceeding its training domain range and entirely new raw water quality and operational conditions. The results of both studies are presented in the following sections.

4.4.1.1. Out-of-range Inputs

Models were developed for both clarifier effluent turbidity and clarifier effluent colour using data from a previous study of enhanced coagulation modelling via ANNs. The difference between the models presented by Stanley *et al.* (2000) and those discussed here is that all four models developed for the current study featured data sets that were truncated according to the value of a key model input variable. The boundaries between the modelling data set and the expanded domain data set for each of the models are presented in Table 4.1.

The model developed for all data patterns that contained raw water turbidity values of 31 NTU or less predicted clarifier effluent turbidity with a mean absolute error of 0.36 NTU and an R^2 value of 0.59 (Table 4.2). As can be seen in Figure 4.1, the model demonstrates excellent predictive capacity when applied to the production data set for all but a few data patterns. The model was then applied to the expanded domain data set where the range of values observed for raw water turbidity ranged from 32 NTU to 1481 NTU. The results of this expanded domain test indicate that absolute prediction errors tend to increase as the value of raw water turbidity increases (Figure 4.2). In essence, the errors increase as the data patterns move further outside the training domain. Furthermore, the relationship between raw water turbidity and absolute prediction error is defined by a logarithmic relationship as shown by Equation 4.1:

$$\text{Absolute prediction error} = ((2.7032)\text{Ln}(\text{Raw Water Turbidity}))-8.8281 \qquad (4.1)$$

This relationship can be fitted to the data with an R^2 value of 0.81, and can be used to identify acceptable model boundaries. If, for example, the maximum desired absolute prediction error is 1.0 NTU, the model is not valid for patterns that contain raw water turbidities that exceed 37.9 NTU. The relationship can also be used to determine the likely absolute prediction error associated with a pattern containing a particular value for raw water turbidity. If, for example, the model were presented with a pattern for which the raw water turbidity was 50 NTU, an absolute prediction error of 1.75 NTU could be expected.

The results for each of the two colour models, when applied to production set data, are presented in Table 4.2. As with the turbidity model, the models show excellent predictive capacities with R^2 values in the range of 0.89 to 0.90 and mean absolute errors in the range of 0.26 to 0.28 TCU. The results for the model where the 75th percentile value of raw water colour was used as the boundary between the modelling and expanded data sets are presented graphically in Figure 4.3. As can be seen in Figure 4.4, a plot of model prediction errors against raw water colour for both the original (0 to 75th percentile) data set and expanded data sets, model prediction errors are tightly distributed around a mean of zero for the original data set. When the model is applied to data where raw water colour is out of bounds however, the model prediction errors begin to increase and diverge. There does appear to be a transition zone, from the boundary between the two data sets up to a raw water colour value of approximately 20 TCU, where model predictions do not appear to be any worse than on the original data set. This suggests that the model can potentially remain valid even when values of raw water colour exceed the maximum value observed in the training domain by approximately 7 TCU. Unlike the logarithmic relationship observed for turbidity model predictions on expanded data (Figure 4.2), the relationship between the value of raw water colour and absolute prediction error appears to be linear (Figure 4.5) and can be expressed by the following equation:

Absolute prediction error =((0.0743)*(Raw Water Colour))-0.3826 (4.2)

This equation accounts for the prediction of absolute error with an R^2 value of 0.49, and can again be used to identify acceptable model boundaries, as previously discussed.

For cases where all data with a raw water colour value at or below the 95th percentile were used in the development of the model, different trends and relationships were obtained. As can be seen in Figure 4.6, the model prediction errors are again tightly distributed around a mean of zero for the original model data. When the model is applied to the expanded domain data however, the vast majority of absolute prediction errors are positive. Since prediction errors are calculated by subtracting the model prediction value from the observed value, this observation suggests that the model is consistently under-predicting the value of clarifier effluent colour in the expanded data domain. As with the aforementoned clarifier effluent colour model, the relationship between raw water colour and absolute error appears to be linear, as demonstrated in Figure 4.7. The equation of the least squares regression line through the data points is as follows:

Absolute prediction error =((0.049)*(Raw Water Colour))-1.458 (4.3)

The data points show more scatter around the regression line than observed for the first colour model and, consequently, a R^2 value of 0.34 was observed. In comparing the equations generated by linear least squares regression through the plots in Figures 4.5 and 4.7, two obvious differences arise. Both the slope of the regression line and the y-intercept are greater for the model where the 75th percentile boundary was applied than for the model where the 95th percentile boundary was applied. When 95% confidence

limits for the estimates of the regression slope and intercept are taken into account however, there is no significant difference between the estimates for the two models. As can be seen in Table 4.5, the 95% confidence interval of the slope and y-intercept estimates for the two models overlap, thereby negating the significance in the difference between the two models' estimates.

The results of both the clarifier effluent turbidity model and the clarifier effluent colour models suggest that the impacts of exceeding the training domain of key process variables when making model predictions are process and variable specific. The turbidity model deteriorated according to a logarithmic relationship when the range of raw water turbidity was expanded, while absolute prediction errors could be related to increases in raw water colour using a linear relationship. These results therefore highlight the need to determine the relative importance of model variables and their effects on model predictions outside the training domain prior to implementing ANN models in process control. The results also suggest that for input variables that are known to heavily influence the value of the output variable, predictions made using values outside of the model training domain should be applied with care. Both the turbidity and colour models deteriorated rapidly when the values of key model input variables were extended beyond their training domain.

4.4.1.2. New Water Quality and Operational Data

The data set used in the development of models to predict clarifier effluent colour, originally presented by Stanley *et al.* (2000) and spanning operations from May 1995 to April 1998, was classified into 7 categories using Kohonen ANNs. The number of categories was selected based on operational knowledge provided by EPCOR Water Services as well as previous modelling experience. Three model input variables, raw water turbidity, raw water colour, and raw water temperature, were used in the classification. The mean values for each of these inputs for each of the seven categories are presented in Table 4.4. The Kohonen ANN is adept at separating the data into distinct categories that match the raw water quality types commonly observed at EPCOR facilities. The water type generally follows seasonal boundaries with separate categories for winter, fall, summer, and transition data. Categories also exist for special events such as spring thaw and summer storm events, which occur infrequently but are extremely important for process operations.

Two separate ANN models were developed for specific categories or combinations of categories. The first was developed for Category 1 data, which represents winter raw water quality conditions when the raw water source is typically under ice cover. Category 1 data patterns occur approximately 43 % of time and are defined by low and stable values of raw water turbidity, colour, and temperature. The second model was developed for summer raw water quality data, as defined by Categories 4, 6, and 7. Raw water quality during the summer months, typically from late May until early September, is

extremely variable, due to the presence of summer storm events. Both models demonstrated good predictive capacity when applied to the production set data, as evidenced by the high coefficients of multiple determination and low mean absolute errors presented in Table 4.5.

The model developed on Category 1 (winter) data was applied to all other categories of data with varying degrees of success, with model results listed in Table 4.6. The model performed reasonably well when applied to data in Categories 2 and 5 with R^2 values of 0.49 and 0.30 and mean absolute errors of 0.68 and 1.23 TCU, respectively. Category 2 data corresponds to transitional (late spring/fall) data, while Category 5 data corresponds to conditions observed during the fall. Based on these results, it appears as though the model trained on winter data performs best when applied to categories that are the most similar; fall and spring data are the closest, in terms of raw water quality, to winter data. The model results when applied to Category 2 data are presented graphically in Figure 4.8. The model appears to loosely follow predominant trends, but rarely makes accurate predictions. The worst prediction performance was observed for Category 3 and Category 4 data, which correspond to spring thaw and summer storm events, respectively. As can be seen in Figure 4.9, the winter model is clearly unable to capture clarifier performance during these infrequent periods of poor water quality. Predictions at peaks are either grossly over predicted or under predicted, and several predictions where clarifier effluent colour is predicted to be negative are made.

As with the Category 1 model, the model jointly developed for Categories 4, 6, and 7 (summer) was applied to the remaining categories. The model results listed in Table 4.7 again demonstrate that model performance is best when applied to data derived from bordering seasons, in this case spring and fall. The results obtained for Category 2 are poor even though the category represents late spring/fall data. The poor performance can likely be explained by the large difference in raw water temperature between Category 2 and Categories 4, 6, and 7.

The results from the two modelling exercises highlight the intuitive importance of ensuring that models are trained on representative data. Where data limitations make it necessary to develop models on a limited data set that may not be fully representative of the full domain of possible raw water quality and operational characteristics, the results suggest that models should be frequently updated as new data become available. Models will continue to offer borderline performance during seasons that are adjacent to the season during which models were developed. Model performance quickly deteriorates, however, for seasons that are not adjacent, making any model-based applications implausible.

4.4.2. Evaluation of Scaling Effects

In order to compare the effects of open scaling, closed scaling, and variable range manipulation on model predictions outside the training domain, the scenario analysis technique discussed in Chapter 3 was employed. For the F.E. Weymouth Filtration Plant

filter effluent particle count model, alum dose was shown to be a key model input variable. In the modelling data set, the range of alum doses observed is from 4.00 to 6.00 mg/L. While there are no operational data outside of this range, it is expected that increasing or decreasing the alum dose beyond these values will impact the process performance. As can be seen in Figure 4.10, models trained using closed input variable scaling generate flat-line predictions outside the training domain. Model predictions for all alum doses less than 4 mg/L are exactly equal in value to those generated for an alum dose of 4 mg/L. Similarly, predictions for alum doses greater than 6 mg/L result in the same value of filter effluent particle counts as a dose of 6 mg/L. Closed scaling essentially eliminates a model's capability to extrapolate beyond the training domain. The open scaling paradigm allows for tentative predictions outside the training domain. As can be seen in Figure 4.10, the slope of the dose response curve gradually decreases as the value of alum dose strays from the range of values on which the model was trained. Since no operational data where alum doses of 1 to 2 mg/L were used exist, it is difficult to verify if the shape of the dose response curve is correct at either extreme. While open scaling definitely has benefits over closed scaling for a variety of problems, it often allows for erroneous result generation. The ODP Plant particle count model was trained with open scaling in Statistica Neural Networks. At the plant, polymer is dosed in the range of 1.6 to 3.9 mg/L. The scenario analysis application developed for the model can generate predictions when applied to data outside this range, even if the data are nonsensical (Figure 4.11). The application suggests that a negative value of filter effluent particle counts will be obtained when a negative dose of polymer is applied; the model can't differentiate between good and bad data. Since, for polymer doses greater than 1.6

mg/L, the value of filter effluent particle counts decreases with added polymer, the model again predicts that a negative filter effluent particle counts will be observed at doses greater than 4.5 mg/L.

An effective balance between open and closed scaling involves the manipulation of the ranges of individual or multiple variables. The F.E. Weymouth Filtration Plant filter effluent particle count model was retrained using closed scaling for all variables except alum dose. The minimum and maximum values for alum dose were manually set to 2 mg/L and 8 mg/L, respectively. As can be seen in Figure 4.10, model predictions are similar to those generated for closed scaling in that they flat line when the dose is outside the new range. By artificially increasing the range however, more meaningful predictions can potentially be made at doses just outside the original 4.00 mg/L to 6.00 mg/L range.

The selection of the appropriate scaling conditions is a balance between extrapolation requirements and alarm-generating capabilities in process control. The closed scaling paradigm is appropriate if the study domain is well defined and data outside the training domain are not expected to be observed. When such is the case, closed scaling has the advantage of ensuring that instrument or other data collection errors will not cause nonsensical model predictions. Open scaling allows for greater flexibility in making predictions using water quality and process data from outside the training domain. This paradigm is particularly useful where models are developed on small or other data sets that are not truly representative of the entire range of conditions that can be expected for a particular process. If open scaling is used, however, adequate alarms must be in place

to ensure that nonsensical input values generated from instrumental or other errors are filtered out when models are used in process control applications. The best compromise between the two scaling options is to manipulate the minimum and maximum values of only those variables where data outside the training domain is likely to be collected. With appropriate alarming, this option protects the utility from the vast majority of nonsensical predictions while allowing greater variations in raw water quality and operational variables.

4.5. CONCLUSION

The determination of model boundaries is an important consideration when applying ANN process models in control applications in the water treatment industry. A number of important training nuances can greatly impact the ability of process models to generate accurate predictions on expanded domain data. Furthermore, the ability of a model to generate predictions on expanded domain data is a function of the distance of a variable's value to the training domain. The relationship between a variable's value and the absolute prediction error is quantifiable, although the relationship appears to be site and variable specific. In combination, these observations suggest that model predictions made on data outside the training domain must be used cautiously; proper setting of alarms is essential to ensure that the effects of prediction errors are maintained within acceptable levels.

4.6. REFERENCES

Stanley, S.J., Baxter, C.W., Zhang, Q., and Shariff, R. 2000. *Process Modelling and Control of Enhanced Coagulation.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO: 167 p.

Table 4.1 Boundaries between modelling and expanded domain data sets

Model output	Input used for classification	Percentile value of boundary	Absolute value of boundary
Clarifier effluent turbidity	Raw water turbidity	75^{th}	31 NTU
Clarifier effluent colour	Raw water colour	75^{th}	13 TCU
Clarifier effluent colour	Raw water colour	95th	35 TCU

Table 4.2 Expanded data domain model results

Model output	Percentile value of boundary	R^2	MAE
Clarifier effluent turbidity	75^{th}	0.59	0.36 NTU
Clarifier effluent colour	75^{th}	0.89	0.26 TCU
Clarifier effluent colour	95th	0.90	0.28 TCU

Table 4.3 95% confidence intervals for the expanded domain colour model regression equations

Model	slope	y-intercept
Clarifier effluent colour (75^{th} percentile boundary)	0.065 to 0.086	-0.71 to -0.10
Clarifier effluent colour (95^{th} percentile boundary)	0.028 to 0.069	-2.48 to -0.43

Table 4.4 Kohonen ANN classification results for clarifier effluent colour model data

Category	Description	% occurrence	Mean Values		
			Turbidity (NTU)	Temperature (°C)	Colour (TCU)
1	winter	43	4.0	0.3	5.0
2	transition (late spring/fall)	6	18.9	4.5	5.7
3	spring thaw	5	142.5	1.1	37.8
4	summer storm	6	262.0	14.3	37.1
5	fall	12	15.9	10.8	8.7
6	early summer normal	16	50.2	16.5	13.7
7	late summer normal	11	20.1	20.2	12.2

Table 4.5 Model results for winter and summer category models

Model	R^2	MAE
Winter (Category 1)	0.76	0.24 TCU
Summer (Categories 4,6,7)	0.84	0.35 TCU

Table 4.6 Category 1 (winter) model results when applied to other categories

Category	Description	R^2	MAE (TCU)
2	transition (late spring/fall)	0.49	0.68
3	spring thaw	0.03	1.70
4	summer storm	0.01	3.13
5	fall	0.30	1.23
6	early summer normal	0.11	1.35
7	late summer normal	0.18	1.53

Table 4.7 Category 4,6, and 7 (summer) model results when applied to other categories

Category	Description	R^2	MAE (TCU)
1	winter	0.05	1.12
2	transition (late spring/fall)	0.06	1.15
3	spring thaw	0.21	3.07
5	fall	0.41	0.48

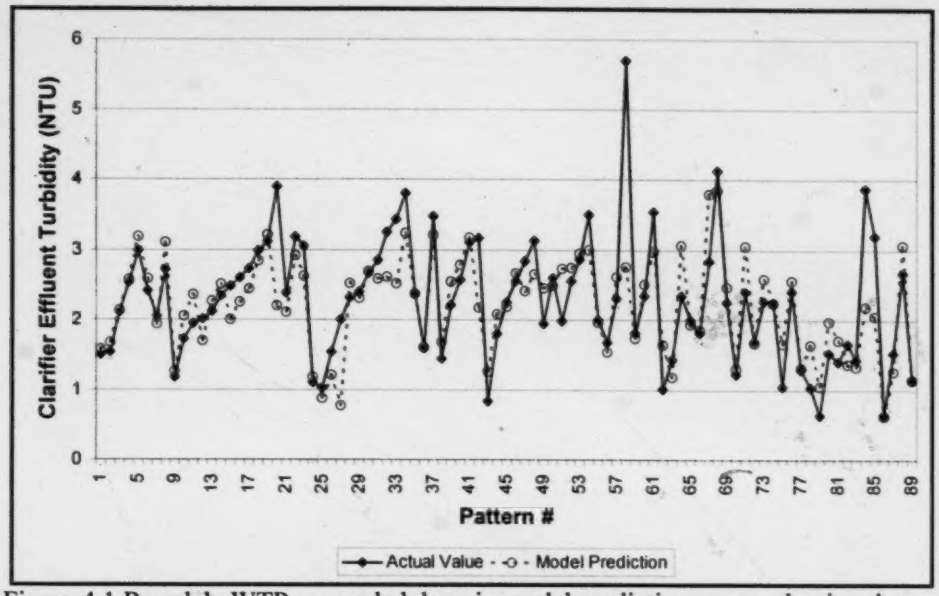


Figure 4.1 Rossdale WTP, expanded domain model predictions on production data set

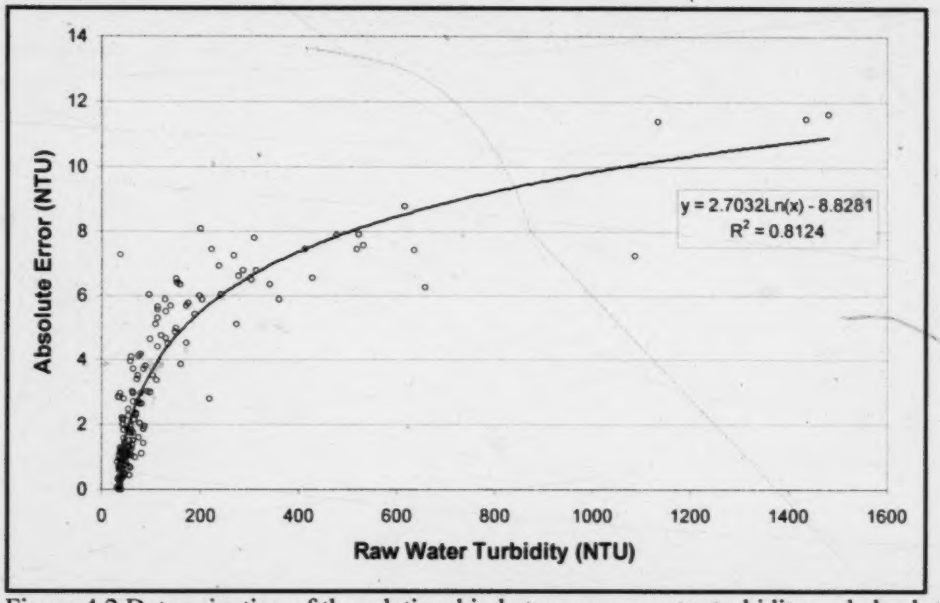


Figure 4.2 Determination of the relationship between raw water turbidity and absolute prediction error

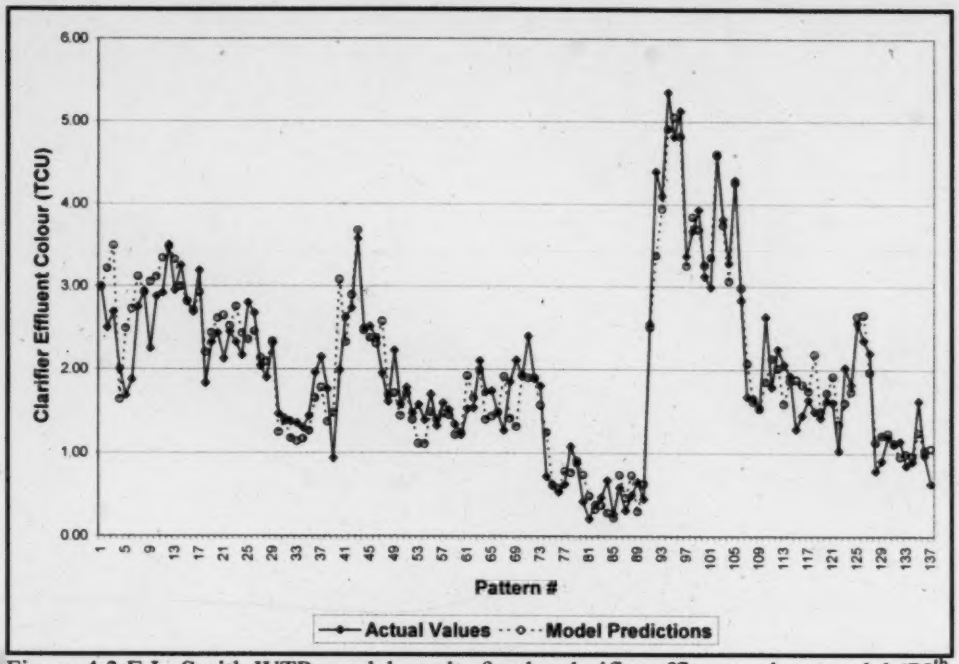


Figure 4.3 E.L. Smith WTP, model results for the clarifier effluent colour model (75th percentile value for raw water colour boundary)

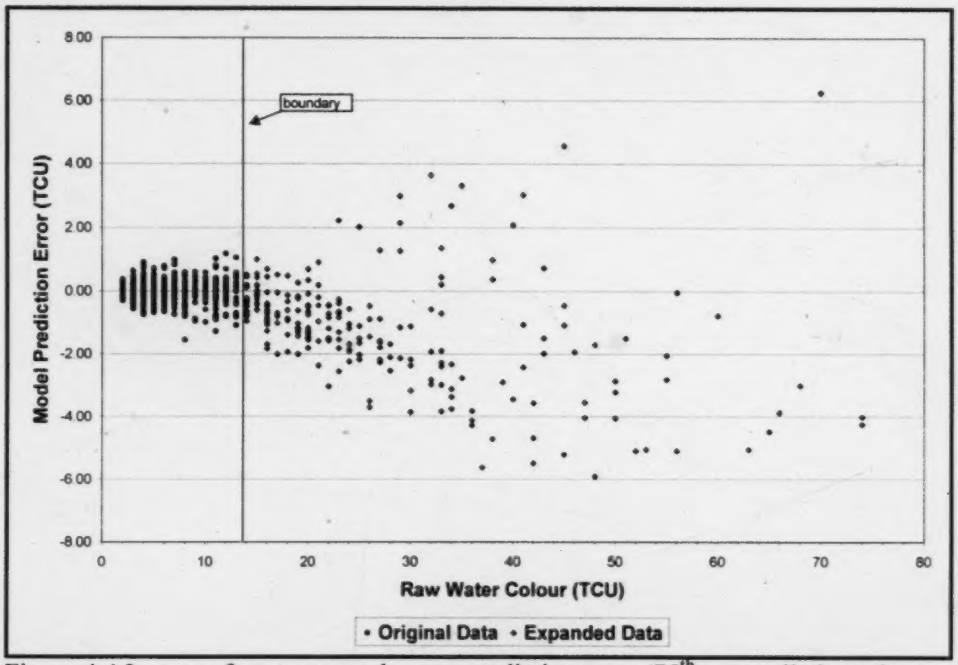


Figure 4.4 Impact of raw water colour on prediction error (75th percentile boundary)

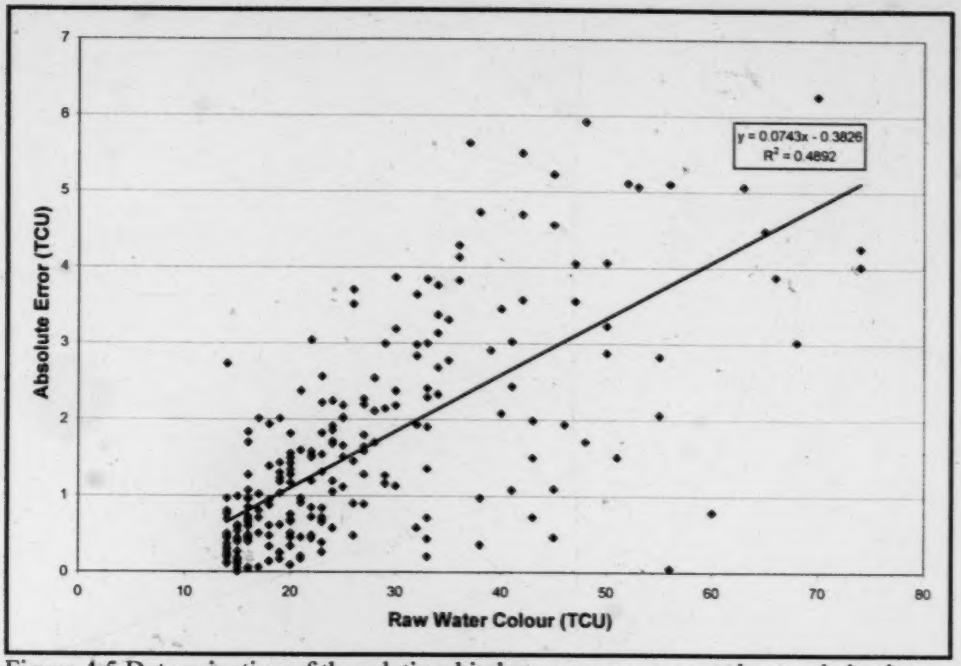


Figure 4.5 Determination of the relationship between raw water colour and absolute prediction error (75th percentile boundary)

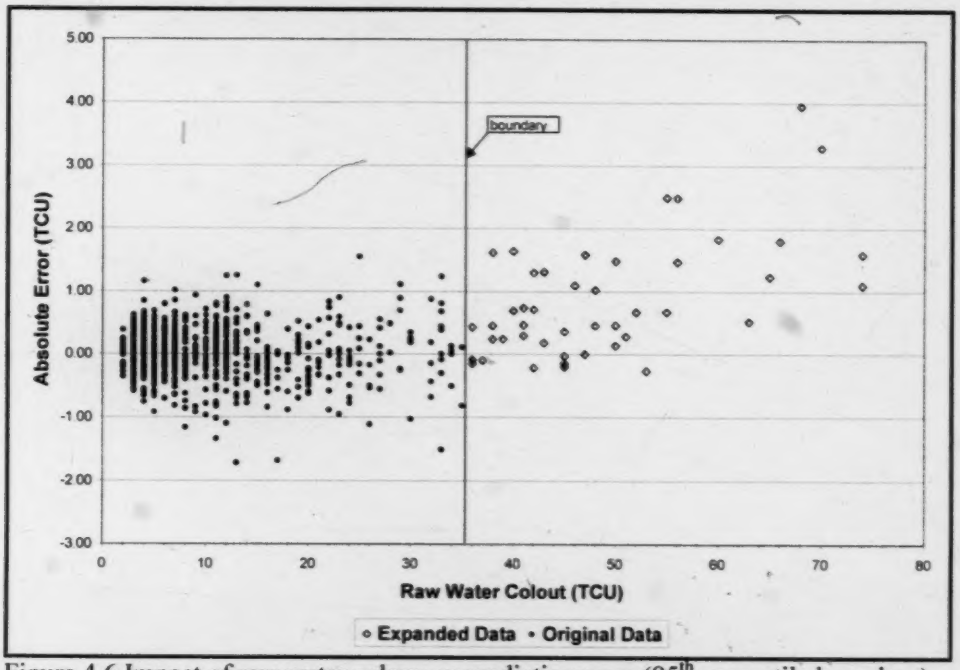


Figure 4.6 Impact of raw water colour on prediction error (95th percentile boundary)

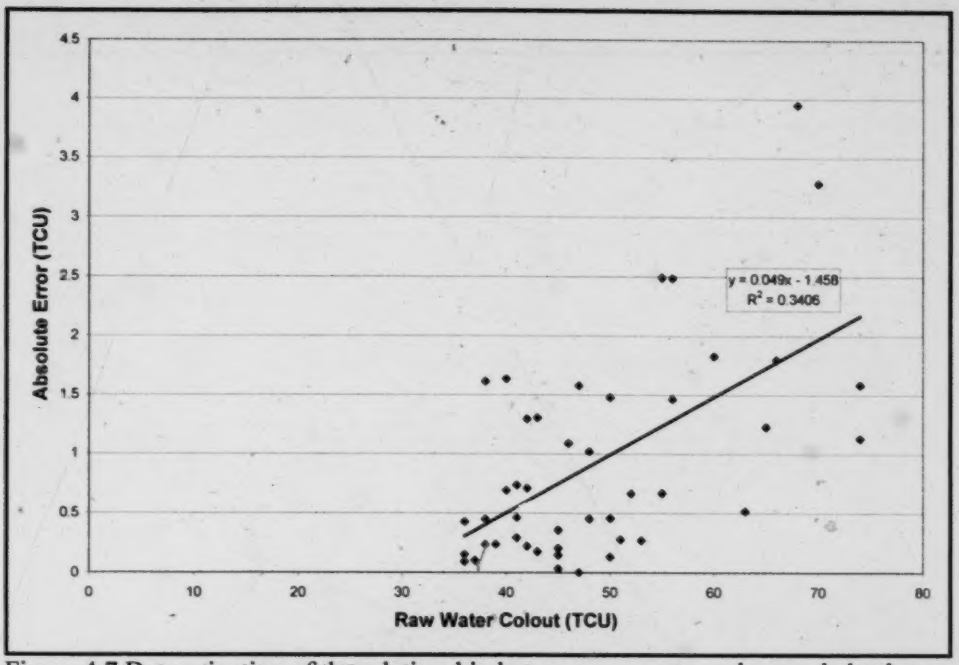


Figure 4.7 Determination of the relationship between raw water colour and absolute prediction error (95th percentile boundary)

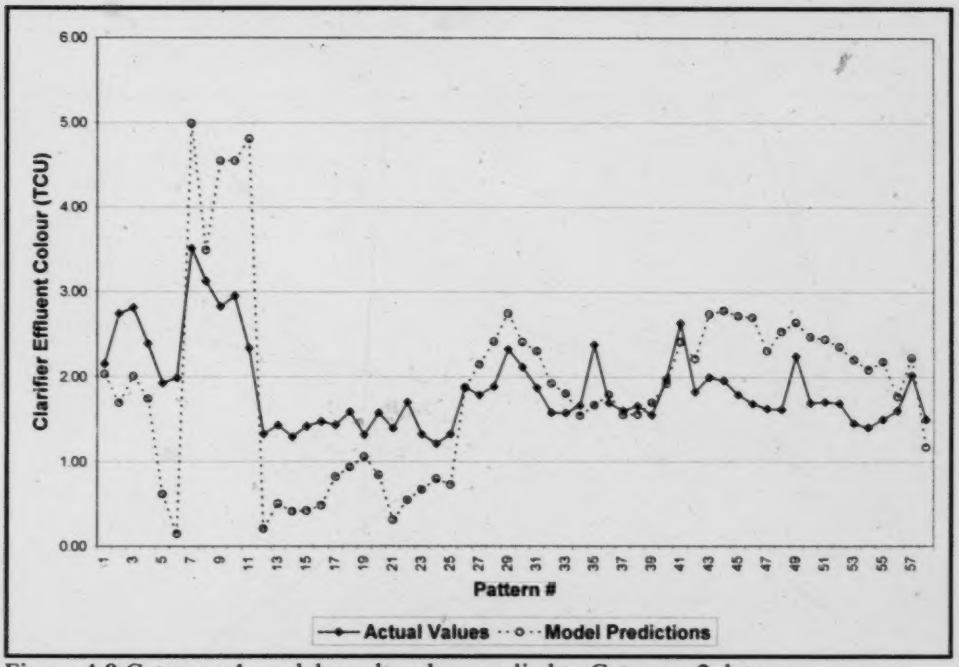


Figure 4.8 Category 1 model results when applied to Category 2 data

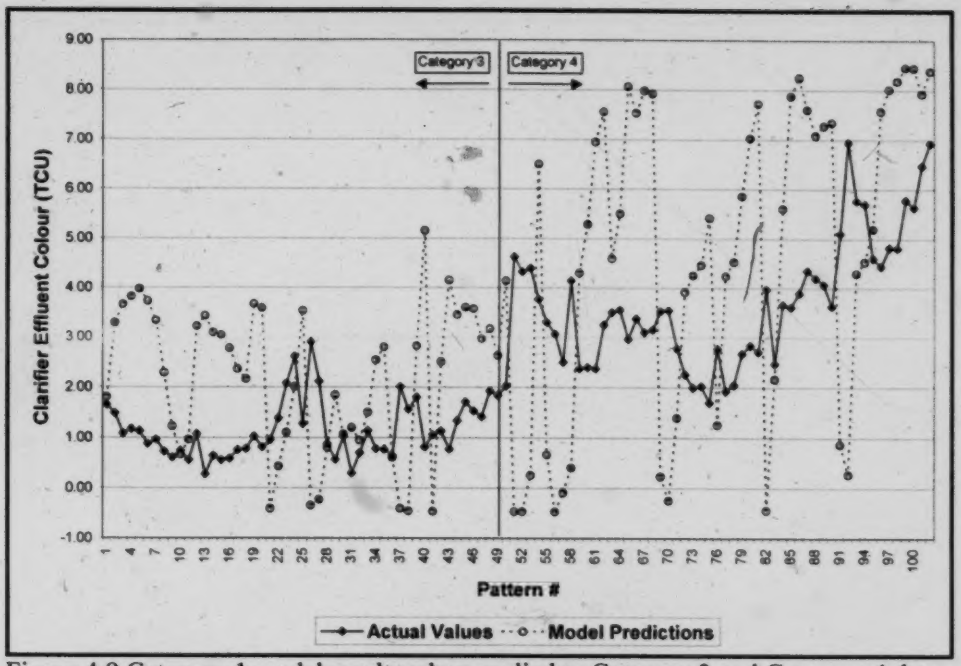


Figure 4.9 Category 1 model results when applied to Category 3 and Category 4 data

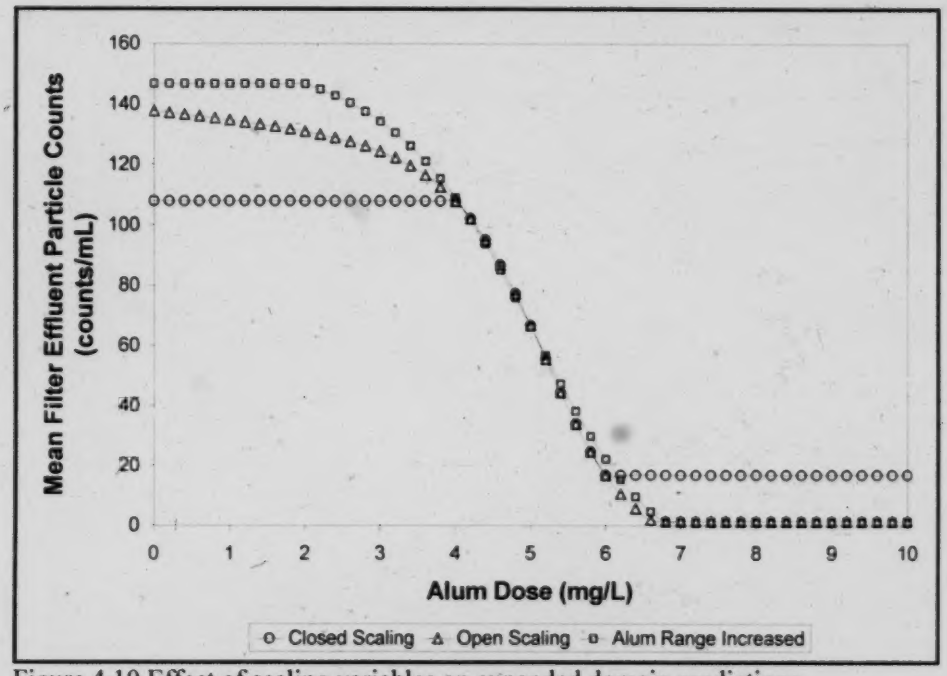


Figure 4.10 Effect of scaling variables on expanded domain predictions

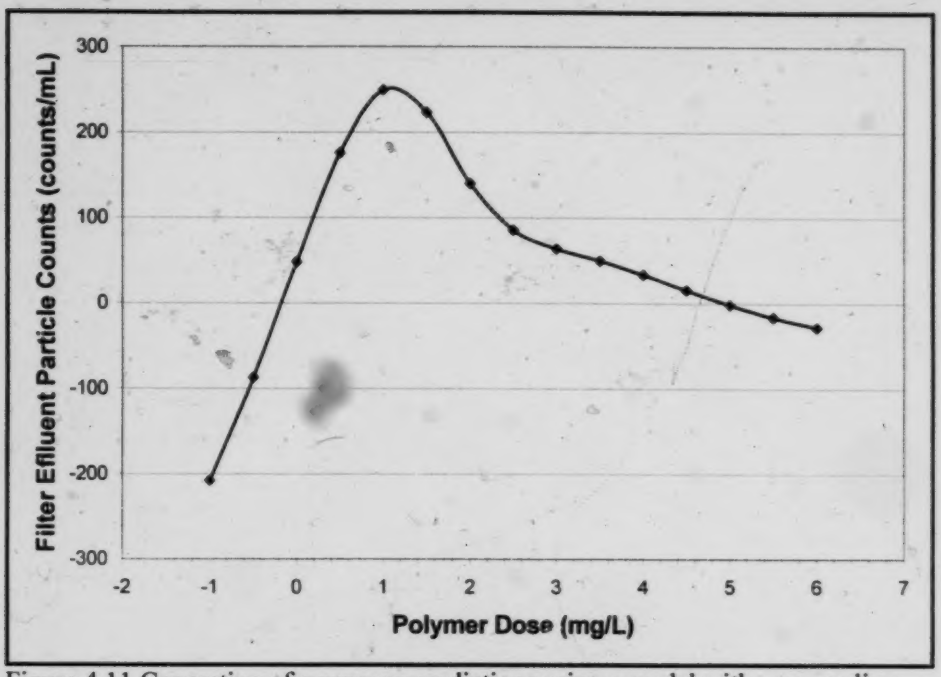


Figure 4.11 Generation of erroneous predictions using a model with open scaling

5. USING ARTIFICIAL NEURAL NETWORKS TO ANALYZE PILOT-SCALE DATA*

5.1. INTRODUCTION

In the water treatment industry, experiments conducted at pilot-scale facilities are often used as the final proving ground for new and modified processes prior to full-scale implementation. The primary goal of pilot testing is to evaluate the effects of varying one or more process parameters on process performance. In many industries where such evaluations are desired, factorial design experimentation has become the favored approach to conducting and analyzing experiments over the past quarter century. The method, popularized by Box, Hunter, and Hunter (1978) allows the experimenter to determine the effects of main factors and their interactions with each other through an optimized series of runs. The design can be augmented through additional runs in order to map a non-linear response surface (Lawson and Erjavec 2001). This feature would suggest that the technique would be well suited to application in the analysis of pilot-scale data in the drinking water treatment industry. Unfortunately, the technique assumes that the effects of the fixed factors under evaluation are not confounded by the effects of random factors. In water treatment pilot testing, it is often impossible to control the values of all the factors that contribute to the outcome of a particular process. When studying the effects of coagulation on particle removal, for example, the experimenter is concerned primarily with the effects of a number of fixed or controllable factors such as

* A version of this chapter has been published. Baxter, C.W., Smith, D.W., and Stanley, S.J. 2001. Using artificial neural networks to analyze pilot-scale data. In *Proceedings of the 1st IWA Conference on Instrumentation, Control, and Automation*. Malmo, Sweden: IWA. 563-569.

coagulant dose and plant flow. Other factors, such as influent temperature, pH, and other raw water characteristics may not be controllable and are therefore considered random factors. Unless the experiments are conducted over a short period of time, the values of the random factors will vary, thereby masking the true effects of the factors under examination.

Where raw water quality variations are present during pilot-scale data collection, analyzing the data and drawing meaningful conclusions from the analysis becomes a formidable challenge. Experimenters often conduct one-variable-at-a-time studies and draw conclusions from simple scatter plots of the resulting data. One of the most common approaches to analyzing pilot-scale data involves the application of empirical modelling techniques. The experimental data are treated as though they were observations sampled from a normally distributed population, and multiple regression analyses are applied.

The purpose of the current study is to demonstrate a simplified method of analyzing pilot-scale data using the artificial neural network (ANN) technology. The main advantage of applying the ANN technique to the analysis of pilot-scale data is its ability to accommodate both fixed and random factors in the analysis with ease, regardless of whether or not nonlinear relationships between factors and the response variable are present. In addition to isolating the effect of each fixed factor, as is the common practice in traditional pilot data analysis, the ANN technique generates a predicted value of the desired output for any reasonable combination of factors, both fixed and random. As

such, the experimenter is able to gain valuable insight into interactions between factors, as well as the contribution of random factors to the value of the output variable.

In order to demonstrate the utility of the ANN technology to pilot-scale data analysis, models were developed for pilot-scale data collected at Metropolitan Water District of Southern California's (MWD's) Oxidation Demonstration Project (ODP) Plant and EPCOR Water Service's pilot plant. The results are compared to those generated using multiple regression analyses, the most common empirical modelling approach for water treatment process data, and assessments of model performance and utility are made.

5.2. THE ANN TECHNOLOGY

The ANN technology is an artificial intelligence technology that attempts to mimic the human brain's problem solving capabilities. ANNs are capable of self-organization and learning; patterns and concepts can be extracted directly from historical data (Baxter, Stanley, and Zhang 1999). When presented with data patterns, sets of input and output data that describe the problem to be modelled, ANNs map the cause-effect relationships between the model inputs and outputs. This mapping of input/output relationships in the ANN model architecture allows developed models to be used to predict the value of the model output variable, given any reasonable combination of model input data, with satisfactory accuracy. Detailed discussions of ANN modelling and applications are presented in Chapters 2 and 3, respectively.

5.3. MULTIPLE REGRESSION ANALYSIS

In modelling of engineered systems, a recurrent problem is one where a response or responses (Y) are known to depend upon a system of variables (x_1, x_2, ..., x_k) and no physical mechanism of describing the dependence is available. According to conventional statistics practice, the effects of changing the levels of the independent variables (x_i) on the response, or dependent, variables can best be determined using multiple regression analyses. Such analyses can be based on fitting either linear or polynomial equations, as thoroughly discussed by Lawson and Erjavec (2001).

5.3.1. General Characteristics of Multiple Regression Analyses

Multiple regression models can be used both to generate predictions based on previously unseen data, and to explain observational relationships. In multiple regression, the general form of the model is:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad (5.1)$$

where β_o is the true Y-intercept, β_i (where i = 1 to k) is the true slope of the regression surface in the x_i direction, and ε is an error term. The regression equation is most easily solved using the method of least squares, which generates estimates for each of the beta (β) coefficients. These estimates are denoted b_i.

In order to quantify the goodness of fit of the regression model, the coefficient of multiple determination (R^2) is applied using the following equation:

$$R^2 = (SST - SSE)/SST \tag{5.2}$$

where SST is the sum of squares (total), and SSE is the sum of squares (error). SST is the sum of squared deviations of the observed values of y from their average, while SSE is the sum of squared deviations of the observed values of y from their model-predicted values. The coefficient of multiple determination, which has a range of 0 to 1 describes the proportion of the variation in Y which is explained by the variation in the set of independent variables. An R^2 value of 1 indicates that 100% of the variation in Y is accounted for by the model.

5.3.2. Assumptions Implicit in Multiple Regression

The multiple regression model is based on the fundamental assumptions that the effects of each independent variable on the dependent variable are linear and additive (Berry and Feldman 1985). While statistical methods for determining whether or not these assumptions hold exist, a better strategy for verifying the assumptions is through a theoretical consideration of the model variables. In water treatment process modelling, for example, both assumptions are known to be routinely violated. Water quality variables often have nonlinear responses, while chemical additions can have

multiplicative effects. Such violations are largely ignored in the application of regression techniques in the drinking water treatment industry, as the proper treatment of said violations requires more sophisticated statistics.

In addition to the assumptions of linear and additive effects, it is assumed that data used in the generation of multiple regression models are free of measurement errors. While error-free observations can rarely be obtained in experimental science, the presence of proper quality assurance and quality control protocols can reduce measurement errors and, consequently, the impact of such errors on regression analyses.

5.3.3. Checking the Adequacy of the Model

The difference between the observed (y_i) and model predicted ($\hat{y}_i$) values of the dependent variable for a given set of independent variable values, specifically $y_i - \hat{y}_i$, is the prediction error or residual. In order for a multiple regression model to be valid, four assumptions concerning the model residuals must be met:

- the residuals are normally distributed;
- the residuals have a constant variance;
- the residuals are independent; and
- the residuals have a mean of zero.

If one or more of the assumptions turns out to be incorrect, the model is not adequate and the data must be re-analyzed with a greater emphasis placed on the cause of the inadequacy.

One key assumption that must be fulfilled in order for a regression model to be valid is that the residuals are normally distributed. In order for this assumption to hold true, a plot of the residuals should look like a random sample from a normal bell-shaped distribution. Histograms and normal scores plots are useful tools for detecting the normalcy of the residuals. In the case of the latter option, a straight-line plot is expected. Deviations from a straight line can help to identify outliers in the observed data. Another assumption made in regression analysis is that all data points are collected with equal precision; residuals will therefore have a common variance. If this assumption holds true, a plot of the model residuals against the model predictions will be free of obvious trends. The third assumption concerning the residuals relates to their independence with respect to time and the values of the model variables. Plots of residuals in time order, as well as plots of residuals against the values of the dependent and independent variables should be free of obvious trends for the assumption of independence to hold true. The final assumption suggests that a good model will have residuals that are scattered around a mean of zero. Simply calculating the mean of the residuals and comparing it to a value of 0 allows for a verification of this assumption.

5.4. METHODS

As previously discussed, the utility of ANNs in pilot-scale data analysis is demonstrated using two separate case studies. The first case study involves a series of 48 data observations obtained from the Metropolitan Water District of Southern California's Oxidation Demonstration Project (ODP) Plant, a large research facility located in La Verne, California, as part of a study of particle removal through coagulation and filtration. The second case study uses a series of 44 data observations from a study of coagulation optimization at EPCOR Water Service's pilot plant located at the E.L. Smith Water Treatment Plant in Edmonton, Alberta, Canada. ANN models were developed using the NeuroShell2 software package from Ward Systems Group, Inc. of Frederick, Maryland. Models were developed according to a modified version of the protocol presented by Baxter *et al.* (2000). The protocol requires that separate test and production data sets be held in reserve for evaluating the model's performance on out-of-set data. When used to analyze pilot-scale data however, the goal of the ANN technique is to extract as much unbiased information as possible from the available observations. As such, all available data were used in model training. The ANN model results were compared to those developed using multiple regression techniques. All multiple regression analyses were performed using the multiple regression module of the Statistica statistics software package from StatSoft of Tulsa, Oklahoma.

5.5. RESULTS AND APPLICATIONS

5.5.1. ODP Plant Analysis

With respect to the ODP plant data, the goal of pilot testing was to determine the effects of various operating conditions on particle removal through filtration. The study used a sound experimental design; three different fixed factors were identified and their effects were studied at multiple levels, as depicted in Table 5.1. The 6 different alum doses, 4 different polymer doses, and two different plant flows identified in Table 5.1 resulted in a total of 48 runs. As can be seen in Table 5.2, the values of five key raw water quality variables known to affect the process varied significantly over the course of the study period. The most variable of these, as measured by the coefficient of variation (COV) were particle counts, turbidity, and temperature. The large variation in raw water quality was attributed to the fact that the study spanned several seasons; data collection occurred from August 22, 1999 to February 23, 2000. When dealing with such variability in random or uncontrolled factors, the traditional approach involves randomizing the order of the runs to reduce confounding the effect of these factors with the effects of the fixed factors being studied. All runs were randomized during data collection in this study.

In evaluating particle removal by filtration, measurements of filter effluent turbidity and filter effluent particle counts are often employed. While particle count measurements are more sensitive to changes in particle concentration than turbidity measurements, government regulations in North America continue to use turbidity measurements to

evaluate particle removal. Two separate models were therefore developed for the ODP Plant data; the variables involved in model development are presented in Table 5.3.

5.5.1.1. Multiple Regression Model Results

The ODP Plant data were subjected to multiple regression analysis using the standard methodologies previously discussed. Stepwise regression, where independent variables are either added to (forward stepwise regression) or subtracted from (backward stepwise regression) the regression model one at a time was also employed. The same 8 variables that were used in ANN model development (Table 5.3) were used as independent variables in the regression model.

Both filter effluent particle counts and filter effluent turbidity were best modelled using standard multiple regression where all independent variables were included. The particle count model had an R^2 value of 0.54 and predicted the model output with a mean absolute error of 306.7 counts/mL. As can be seen in Figure 5.1, model predictions rarely matched the observed values for the dependent variable, resulting in poor trending and peak prediction. The large mean absolute error resulted in cases where the value of filter effluent particle counts was predicted to have a negative value, a physical impossibility. The turbidity model offered slightly better results with an R^2 value of 0.61 and a mean absolute error of 0.09 NTU. As can be seen in Figure 5.2, however, model predictions are still inadequate with poor peak prediction and several negative predictions.

The beta coefficients and the slope estimates (b_i) of the regression models, along with their standard errors at the 0.05 level of significance are presented in Table 5.4. The beta coefficients are the slope estimates that would have been obtained had the independent variables been standardized to a mean of 0 and a standard deviation of 1 prior to regression. The advantage of beta coefficients is that their magnitude allows for the comparison of the relative contribution of each independent variable in the prediction of the dependent variable. Of the slope estimates, only those of alum dose and percent State Project Water were found to be statistically significant for the particle counts models. Similarly, only alum dose, polymer dose, and percent State Project Water had significant coefficients for the turbidity model.

Before further investigation of model results, the regression models' adequacy needs to be verified. Normal plots of the residuals resulted in straight-line plots for both models. For the turbidity model, the presence of a prediction outlier was detected through the normal plot, as can be seen in Figure 5.3. Plots of model residuals against model predictions for both the particle counts and turbidity models showed a slight diverging trend. This suggests that the assumption of constant variance does not hold as the absolute values of the residuals increase with model predictions. Plots of the model residuals against each of the independent variables in the models were free of trends. When plotted against the dependent variables however, residuals in both models increased as the value of the models' dependent variables increased (Figure 5.4). Finally, the mean value of the residuals for each of the models was calculated to be 0.

The poor performance of the regression models can best be explained by the size of the data set and the high degree of multicollinearity among independent variables. In general, multiple regression models perform best when the number of observations is at least ten to twenty times higher than the number of model variables. With 8 independent and one dependent variable in each model, 90 to 180 observations would be required to satisfy this criterion. Given that the current study encompasses only 48 observations, the estimates of the regression line are probably very unstable and unlikely to replicate if the study were repeated. Regarding multicollinearity, which describes the situation where an independent variable is linearly correlated to another or a combination of others, the standard errors of the regression coefficients increase as the degree of multicollinearity increases. Multicollinearity can be detected by regressing each of the independent variables on the set of remaining independent variables. For the particle counts model, plant flow, plant influent temperature, plant influent turbidity, and percent State Project Water each had R^2 values that exceeded 0.53, the value for the complete model, when regressed against the other variables. In particular, when the remaining independent variables were used as predictors for plant influent temperature, an R^2 value of 0.86 was obtained. In this example, the beta values for plant flow, plant influent turbidity, and percent State Project Water were found to be significant. As such, much of the variability in plant influent temperature can be described by variability in other independent parameters. Similarly, for the turbidity model, plant flow, plant influent temperature, plant influent turbidity, and plant influent dissolved oxygen demonstrated a high degree of multicollinearity with other independent variables. In combination, the small size of the modelling data set as well as the high degree of observed multicollinearity among

independent variables served to increase the standard errors of the model coefficients and decrease model reliability.

5.5.1.2. ANN Model Results

Both the particle counts and turbidity ANN models demonstrated excellent predictive capacity with R^2 values of 1.00 and 0.99, and mean absolute errors of 6.9 ((counts>2μm)/mL) and 0.01 NTU for the particle counts and turbidity models, respectively. The particle count model results are depicted graphically in Figure 5.5, while those for the turbidity model can be found in Figure 5.6. As previously discussed, all 48 data patterns were used in model training. As such, the ANN models are able to generate near-perfect mapping of the relationships between model inputs and model outputs. This approach is not typically used when the purpose of model development is to generate predictions for new data, as the resulting models tend to have poor prediction capacity on new data. For data collected under the auspices of a sound experimental design however, the approach leads to models that interpolate well within the limits of the design. The results suggest that the ANN technique is more suitable to modelling small pilot-scale data sets than the multiple regression approach, as the results obtained through ANN modelling are far superior to those obtained by multiple regression for both filter effluent particle counts and filter effluent turbidity. This argument is further supported by the ability of ANN models to generate predictions across a varied domain of data, as discussed in the model applications section.

5.5.1.3. ANN Model Applications

The completed models can be used to generate important information regarding the relative importance of model inputs, regardless of whether the inputs were fixed or random factors in the experimental design. A combined model-based sensitivity analysis performed on each of the model variables indicated that the most important factors influencing the value of the two output variables were alum dose, polymer dose, % State Project Water, and influent temperature. The sensitivity analysis does not yield information regarding the absolute effect of each factor, but does highlight the need to take into consideration the effect of at least some of the random factors, including raw water temperature, when analyzing the pilot data. The results of the sensitivity analysis also serve as a validation of the selection of fixed factors and their associated levels in the experimental design. Since the ANN models do not appear to be sensitive to the effects of plant flow, for example, it is possible to conclude that the levels evaluated for the given factor were spread over too narrow a range for a noticeable effect to be observed.

The models can be used to make predictions within the limits of the experimental design that yield more information concerning process performance than traditional statistical analyses. The values of one or more of the fixed or random factors can be altered alone or in combination, and the resulting effect on the model output variable can be determined. This type of analysis is particularly useful when the results are presented graphically. Figure 5.7 was generated using the filter effluent turbidity model by varying the alum dose and polymer dose within the limits of the experimental design while holding the

remaining variables constant at their mean values. The figure suggests that increasing both alum dose and polymer dose can assist in turbidity removal, an observation supported by process knowledge at the facility. The model also suggests there are many opportunities for overdosing, as much of the surface at higher dosing levels is flat. This information can be used in designing further pilot scale experiments or in planning operation strategies for full-scale operations. As previously discussed, a sensitivity analysis on the model variables revealed that influent temperature is a key factor in determining process performance. As can be seen in Figure 5.8, filter performance tends to decrease with increases in temperature, as well as increases in percent State Project Water. As it is not possible to control the temperature or the quality of the facility's raw water, this observation suggests that due consideration must be given to water quality fluctuations when planning pilot-scale tests that span several seasons.

5.5.2. EPCOR Water Services Pilot Plant Analysis

In 1994, EPCOR Water Services undertook a comprehensive study to evaluate a number of different primary coagulants and coagulant aids for use in their two full-scale facilities in Edmonton, Alberta, Canada. As is the norm in the water treatment industry, the study was conducted using bench and pilot-scale experiments, the results of which were extrapolated to full-scale operations. Unlike the ODP Plant study however, the pilot-scale tests did not follow an organized experimental design. Instead, the experimenters took advantage of the multiple treatment trains at the pilot plant to conduct simultaneous comparative analyses of candidate coagulants.

The data selected for analysis by the ANN technique encompasses approximately one-half of the study data. The commonality between the 44 data patterns selected for analysis is the use of alum as the primary coagulant along with an anionic polymer coagulant aid. A complete listing of the variables measured during the study, along with a statistical analysis of the study data, is presented in Table 5.5. Many of the influent variables demonstrate a high degree of variability. As with the ODP study, the pilot plant study spanned several seasons. The source water for the pilot plant is under ice cover for 5 months per year. As the river ice thaws, raw water quality deteriorates significantly, thereby explaining the variability in variables such as turbidity and colour. During the study from which the modelling data set was obtained, alum was not dosed according to an experimental plan. Instead, the alum dose was selected by mirroring dose changes in EPCOR's full-scale facilities. With few exceptions, the polymer dose was held constant at a value of 0.26 mg/L during the study. The model output for the EPCOR pilot plant study was clarifier effluent turbidity.

5.5.2.1. Multiple Regression Model Results

A multiple regression model was built for the pilot plant data using the methodology previously discussed. When all 7 model input parameters were used in the model, an R^2 value of 0.35 and a mean absolute error of 0.34 NTU were obtained. A plot of the model results is presented in Figure 5.9. In spite of the small coefficient of determination, the model appears to follow general trends in the observed data better than the regression

models obtained for the ODP and had no negative prediction values. Regarding the analysis of model coefficients, of all the independent variables, only influent turbidity was found to have a significant slope coefficient. In addition to the standard multiple regression methodology, both forward and backward stepwise regression techniques were applied. The results obtained from these analyses were inferior to those obtained via standard multiple regression. In fact, when the backwards approach was applied, none of the model variables were found to have significant slope coefficients, and a coefficient of multiple determination of 0.00 was obtained.

Regarding the verification of model adequacy, a normal plot of the residuals was found to be linear, and the mean of the residuals was calculated to be 0, thereby verifying the first and fourth model assumptions. Plots of the residuals against each of the independent variables yielded no observable trends, however, a plot of the model residuals against the observed values of the dependent variable showed an increasing trend, as was observed for the ODP plant regression models. Similarly, a plot of the residuals against model predictions showed a slight diverging trend. These last two observations suggest that the second and third model assumptions, those of constant variance and residual independence, cannot be verified.

The mediocre model performance can again be explained by the size of the data set, the lack of important process information, and the high degree of multicollinearity among independent model variables. Unlike the ODP Plant study, where the values of the controllable factors were set at multiple levels and the runs randomized prior to

experimentation, the values of the controllable factors at the pilot plant were set to match full-scale operations at the time of the study. As such, the data domain is less-organized than that of the ODP study, resulting in a less uniform distribution of data from the controllable factors across the study space. In addition, plant flow was not among the variables measured in the pilot plant study, even though its value is known to impact process performance. With regards to multicollinearity, every independent variable other than polymer dose could be predicted with a higher coefficient of multiple determination, when the remaining independent variables were used as predictors, than that obtained for the original model. In particular, both influent pH and influent temperature could be predicted with R^2 values that exceeded 0.95, which suggests a definite linear correlation among independent model variables.

5.5.2.2. ANN Model Results

An ANN model was developed to predict the value of clarifier effluent turbidity given the values of all the other variables presented in Table 5.5. An R^2 value of 0.96 was obtained when model predictions were compared to actual values, and the mean absolute error of prediction was found to be 0.08 NTU. The model results are depicted graphically in Figure 5.10. As with the ODP study, the ANN model results are far superior, both statistically and graphically, to those obtained by regression analysis. A model-based sensitivity analysis was performed on the model variables, and the values of alum dose, influent turbidity, and influent colour were found to be the most influential in determining the value of clarifier effluent turbidity.

5.5.2.3. ANN Model Applications

As with the models generated for the ODP Plant, the ANN models developed for the EPCOR pilot plant can be used to generate response surfaces for the model output variable. The plot in Figure 5.11 was generated using model predictions for typical spring break-up water, which is characterized by high values of influent turbidity and colour. The plot, which consists of a narrow trough and two peaks suggests that the optimal alum dose depends highly on the value of influent turbidity and that dosing beyond the optimum results in increases in clarifier effluent turbidity.

The impact of raw water quality, as measured by influent colour and turbidity, on clarified water turbidity is presented in Figure 5.12. The plot was generated for typical summer operations that are characterized by elevated values of alkalinity, turbidity, and temperature. As can be seen in the plot, increases in plant influent turbidity can have a tremendous impact on clarified water turbidity if the alum dose, which was held constant at 75 mg/L in the data used to generate the plot, is not optimized. With respect to colour, there appears to be no discernable impact of changes in plant influent colour on clarifier effluent turbidity, even though increases in colour are generally indicative of poorer quality source water. This observation can be explained by the fact that only a limited range of colour values, from 7.3 to 9.9 TCU, were observed over the course of this small pilot-scale study during summer months. The range of values evaluated in Figure 5.12 does not reflect a typical summer range of values, but expanding the range beyond these

values would force the model to extrapolate beyond its training domain. It is important to note that the relationships depicted in Figures 5.11 and 5.12 apply to specific raw water quality conditions observed at EPCOR facilities and are defined by tight boundaries for each of the measured factors. The effects of alum dose and other factors under different raw water quality conditions can be determined using similar graphical analyses.

5.6. LIMITATIONS OF THE TECHNIQUE

As with any method of data analysis, there are several limitations to the ANN technique of pilot-scale data analysis of which potential users should be aware. Many ANN software packages will generate predictions for any combination of values entered for the model input variables, even if such combinations are nonsensical. At EPCOR facilities, for example, it is not possible to have a combination of warm raw water (> 15 °C) and low raw water turbidity (<5 NTU) due to the nature of the facilities' source water. Predictions made on such combinations will be based on extrapolations beyond the model's training domain. As such, users should only apply trained models to data combinations that are known to be plausible and that fall within the study domain. As well, with small data sets, such as those encountered in pilot testing, every data point contributes more information to the model, on a relative scale, than data points in larger data sets. A single erroneous data entry can compromise the integrity of the model. Users should therefore ensure that appropriate quality assurance and quality control mechanisms are in place when collecting data for analysis. In addition, while the technique is invaluable in the analysis of complex nonlinear processes where interactions

between various key factors are poorly understood, sound mechanistic models can often yield more accurate information when such models exist.

Due to the size of the data sets typically encountered in pilot-data analysis, independent evaluations of model performance are generally not possible. Where additional confirmatory data can be collected, however, model performance can be validated. Providing that the models are only applied to data within the study domain, as previously discussed, sound applications can still be developed.

5.7. CONCLUSIONS

The use of the ANN technique in the development of full-scale models and process control applications is currently on the rise in the drinking water treatment industry. These models and applications have typically been developed on large, carefully selected historical data from treatment facilities. The results of the current study suggest that the ANN technique also has application potential in the analysis of pilot-scale data where, due to financial and time considerations, relatively few observations are available. The technique can accommodate both fixed and random factors, thereby facilitating both the design of pilot-scale experiments and the subsequent data analysis. The results of the current study suggest that the ANN technique can provide vastly superior results when compared to multiple regression modelling, the most widely applied empirical modelling technique for multivariate continuous data, for pilot-scale data analysis.

The benefits of the ANN technique do not, however, negate the need for careful planning of pilot-scale experiments. Sound experimental designs that cover a wide range of values for key factors, such as the one presented in the MWD case study, ensure that the models are based on the best and most representative data available. Predictions made from such models will generally lead to more useful information than predictions made from models of happenstance data.

5.8. REFERENCES

Baxter, C.W., Stanley, S.J., Zhang, Q. and Smith, D.W. (2000). Developing artificial neural network process models: a guide for drinking water utilities. In *Proceedings, 2000 Annual Conference of the Canadian Society of Civil Engineering.* London, ON.: CSCE.

Baxter, C.W., Stanley, S.J., and Zhang, Q. (1999). Applications of full-scale artificial neural network models of enhanced coagulation. In *Proceedings, 1999 Information Management and Technology Conference.* New Orleans, LA.: AWWA.

Berry, W.D., and Feldman, S. 1985. *Multiple Regression in Practice*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-001. Beverly Hills and London: Sage Publications. 95 p.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons. New York, NY: 653 p.

Lawson, J., and Erjavec, J. 2001. *Modern Statistics for Engineering and Quality Improvement*. Duxbury, Thomson Learning. Pacific Grove, CA. 810 p.

Table 5.1 Levels of fixed factors investigated during the ODP study

Fixed Factor	Levels Studied
Alum Dose (mg/L)	0.0, 1.0, 3.0, 5.0, 7.0, 10.0
Polymer Dose (mg/L)	0.0, 1.0, 2.5, 5.0
Plant Flow (MGD)[1]	2.0, 4.0

[1] Note: 1 MGD = 3.7854 ML/d

Table 5.2 Statistical summary of raw water quality variables during the ODP Plant study

Variable	Mean	Std. Dev.	COV (%)	Percentile 5th	50th	95th
Temperature (°C)	18.09	5.23	28.92	11.44	19.15	24.13
pH	8.29	0.06	0.74	8.20	8.30	8.40
% State Project Water (% v/v)	26.51	6.25	23.56	22.00	25.50	30.93
Turbidity (NTU)	1.13	0.36	31.64	0.62	1.13	1.82
Particle Counts ((>2μm) /mL)	4130.75	1690.23	40.92	1763.20	4095.00	6545.15
Dissolved Oxygen (mg/L)	9.47	1.52	16.09	7.54	9.30	11.87

Table 5.3 ODP Plant model variables

Filter Effluent Particle Counts Model	Filter Effluent Turbidity Model
Alum Dose (mg/L)	Alum Dose (mg/L)
Polymer Dose (mg/L)	Polymer Dose (mg/L)
Plant Flow (MGD)	Plant Flow (MGD)
Influent Temperature (°C)	Influent Temperature (°C)
Influent pH	Influent pH
% State Project Water (% v/v)	% State Project Water (% v/v)
Influent Turbidity (NTU)	Influent Turbidity (NTU)
Influent Particle Counts ((>2μm) /mL)	Dissolved Oxygen (mg/L)

Table 5.4 ODP Plant multiple regression model coefficients

	Particle counts model				Turbidity Model			
	Beta	std. error	*b*	std. error	Beta	std. error	*b*	std. error
Intercept			-15750.37	11093.01			-3.79	3.66
Alum dose (mg/L)	-0.45	0.12	-74.14	19.40	-0.44	0.10	-0.03	0.01
Polumer dose (mg/L)	-0.20	0.11	-60.76	34.13	-0.32	0.10	-0.03	0.01
Plant flow (ML/d)	0.34	0.27	192.69	154.60	0.21	0.24	0.04	0.05
Plant inf. temperature (°C)	0.50	0.29	54.26	31.91	-0.02	0.36	0.00	0.01
Plant inf. turbidity (NTU)	-0.10	0.24	-131.96	305.72	-0.21	0.18	-0.10	0.08
Plant inf. pH	0.18	0.14	1648.24	1294.57	0.15	0.13	0.51	0.44
State Project Water (%v/v)	0.62	0.16	56.88	14.77	0.65	0.15	0.02	0.00
Plant inf. particle counts ((counts >2 um)/mL)	-0.04	0.21	-0.01	0.07				
Plant inf. dissolved oxygen (mg/L)					-0.51	0.30	-0.07	0.04

Table 5.5 Statistical summary of data collected during the EPCOR Pilot Plant study

				Percentile		
Variable	Mean	Std. Dev.	COV	5th	50th	95th
Influent Turbidity (NTU)	20.49	16.23	79.23	4.94	13.49	52.41
Influent Colour (TCU)	19.96	12.45	62.37	3.50	21.00	40.00
Influent pH	8.03	0.16	2.00	7.89	7.96	8.31
Influent Alkalinity	125.89	6.67	5.30	113.30	126.00	134.85
Influent Temperature (°C)	4.52	6.41	141.82	0.50	1.00	15.50
Alum Dose (mg/L)	71.08	34.51	48.55	19.40	77.80	129.80
Polymer Dose (mg/L)	0.27	0.04	14.81	0.25	0.26	0.39
Clarifier Effluent Turbidity (NTU)	1.05	0.55	52.53	0.40	0.96	1.95

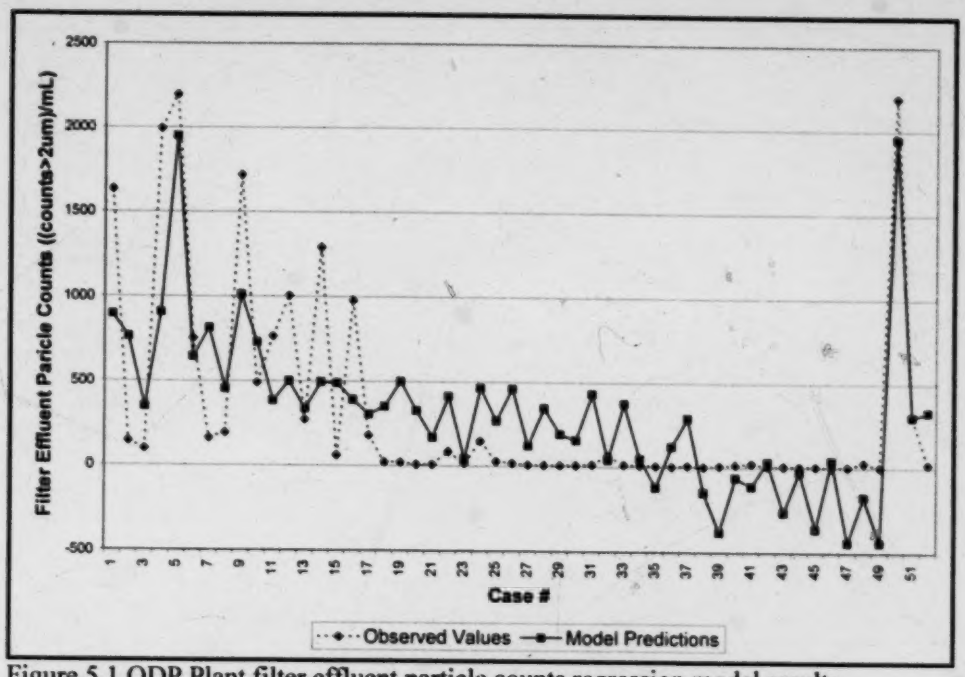


Figure 5.1 ODP Plant filter effluent particle counts regression model results

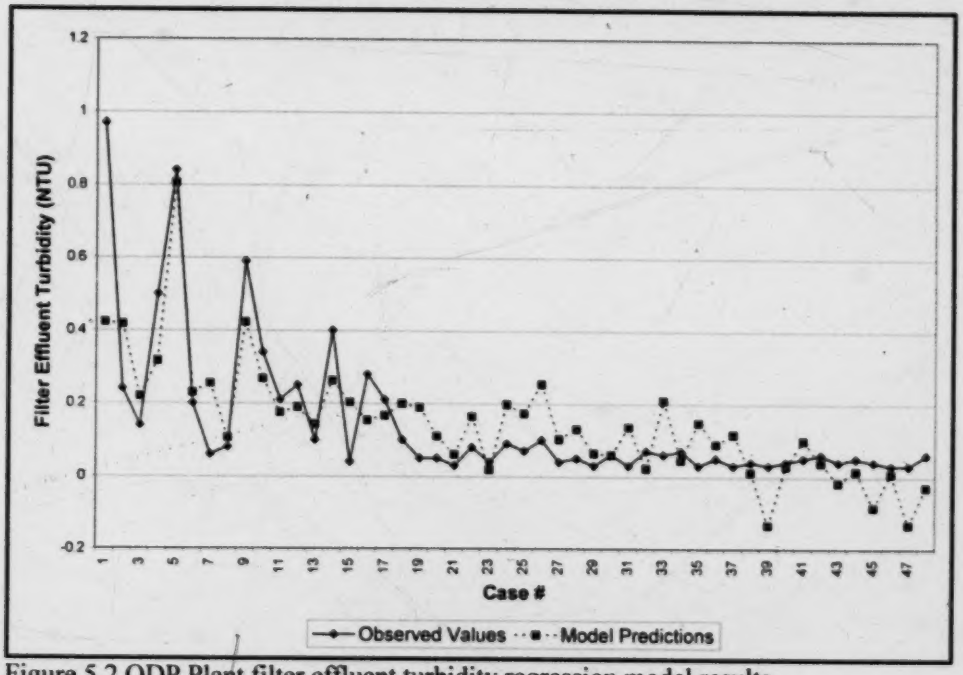


Figure 5.2 ODP Plant filter effluent turbidity regression model results

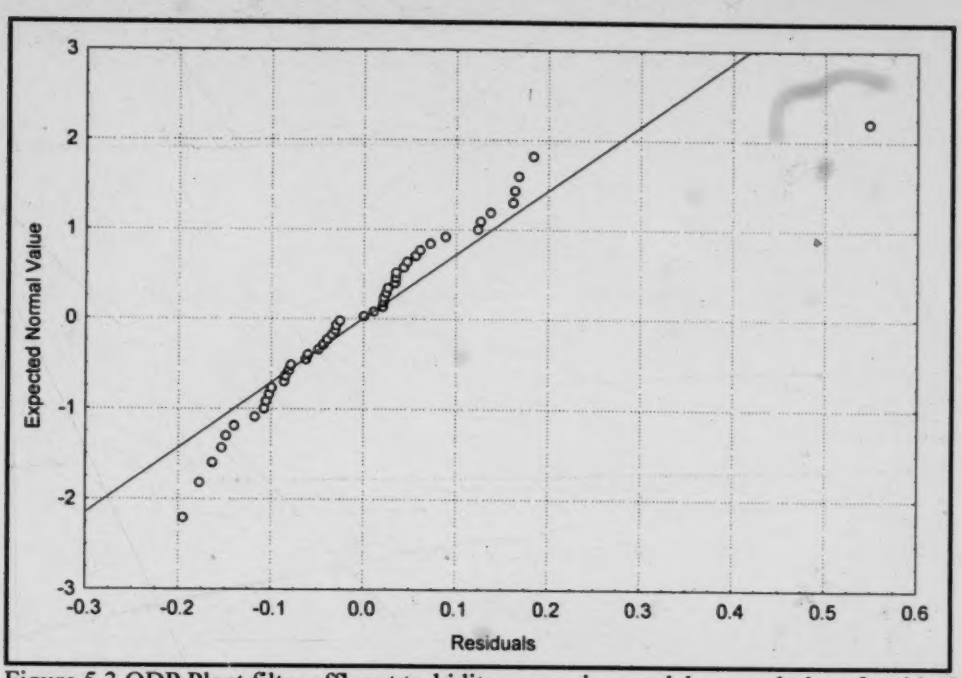


Figure 5.3 ODP Plant filter effluent turbidity regression model, normal plot of residuals

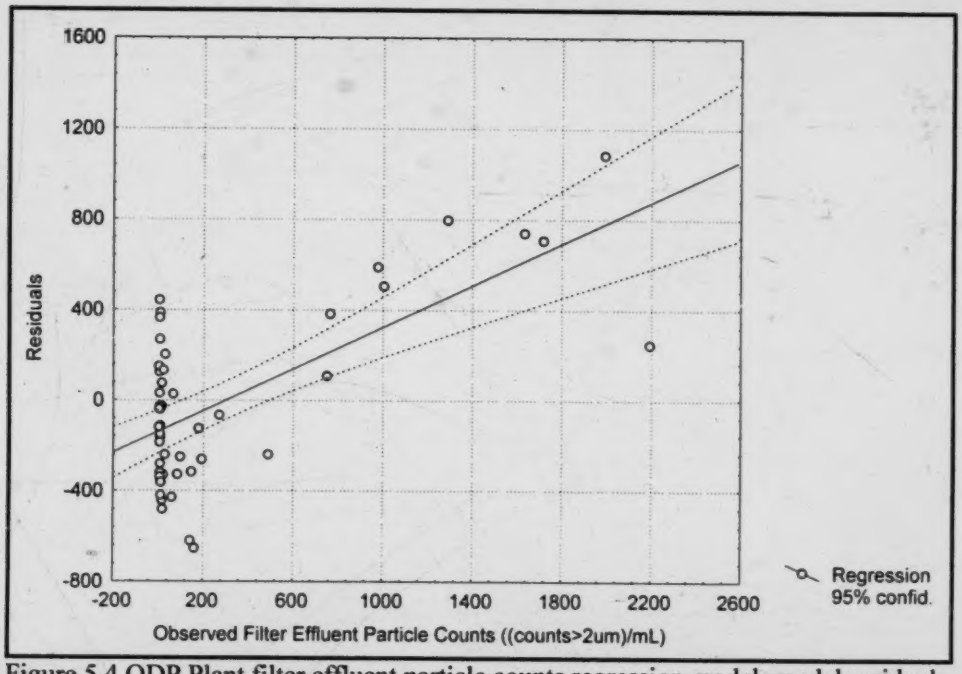


Figure 5.4 ODP Plant filter effluent particle counts regression model, model residuals against observed values

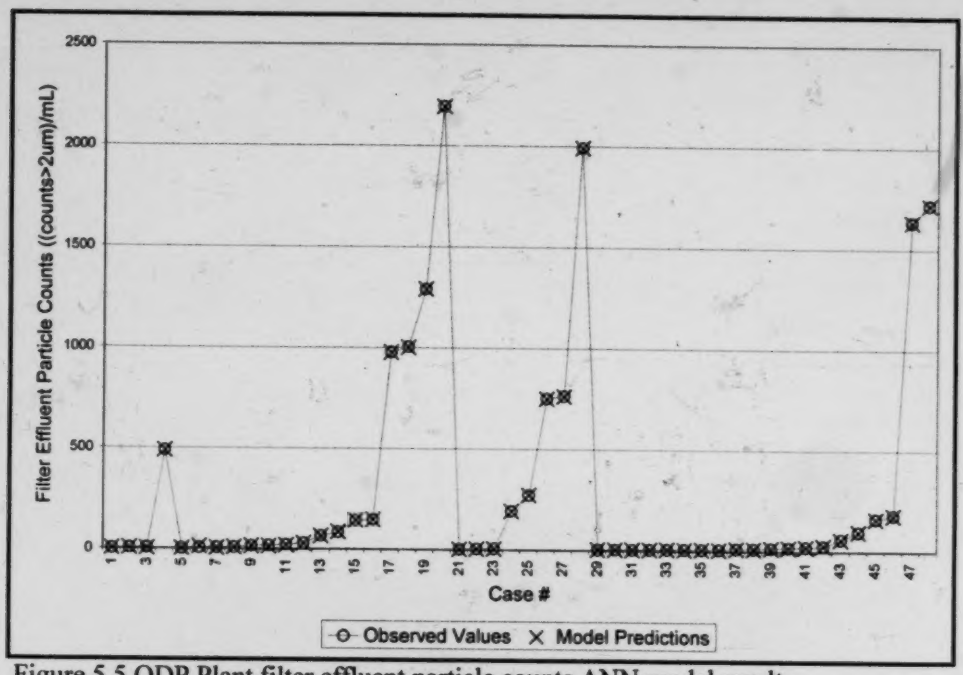


Figure 5.5 ODP Plant filter effluent particle counts ANN model results

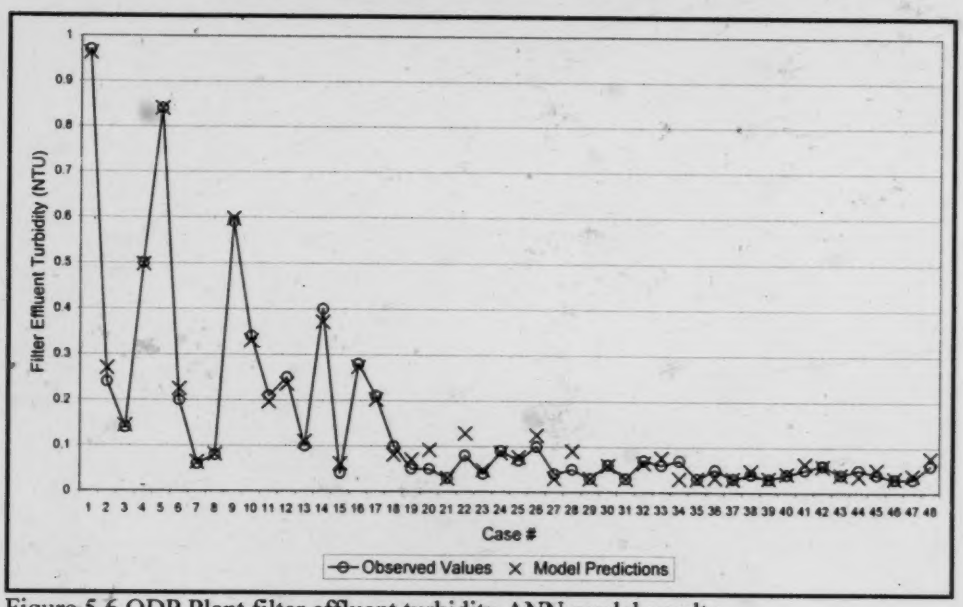


Figure 5.6 ODP Plant filter effluent turbidity ANN model results

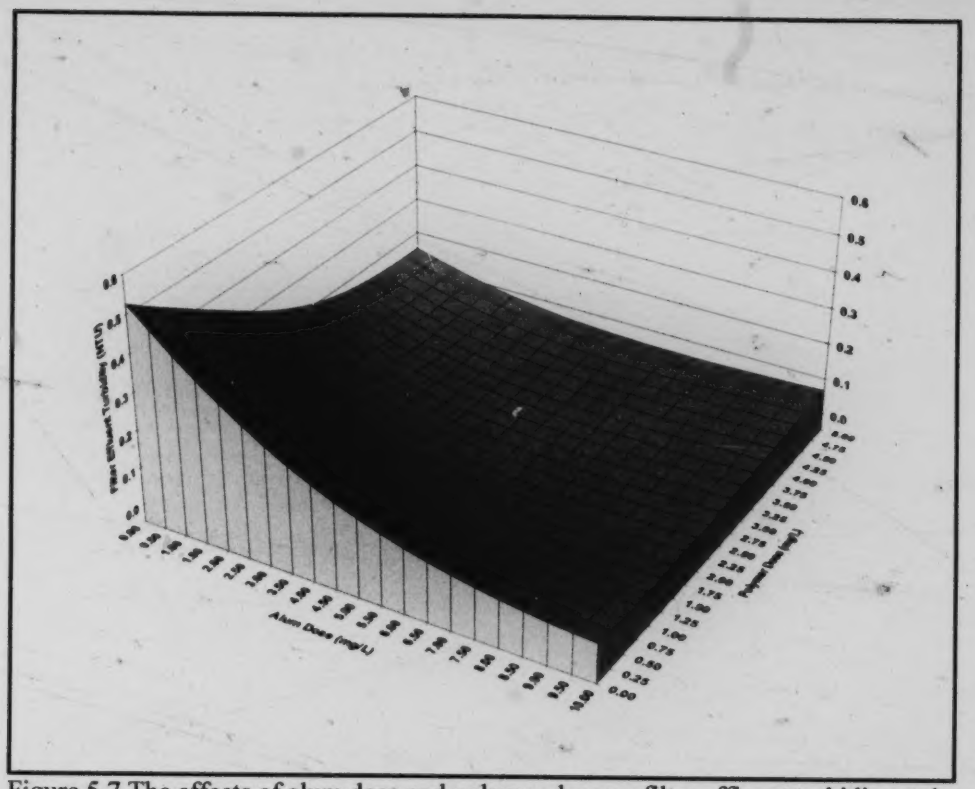


Figure 5.7 The effects of alum dose and polymer dose on filter effluent turbidity at the ODP Plant

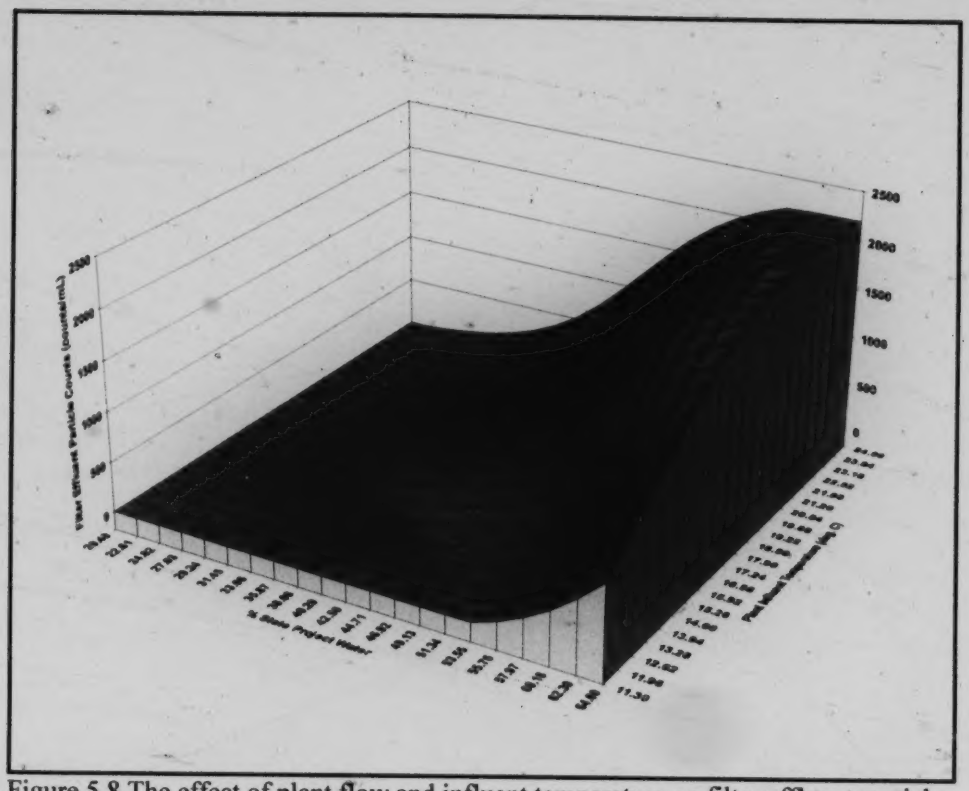


Figure 5.8 The effect of plant flow and influent temperature on filter effluent particle counts at the ODP Plant

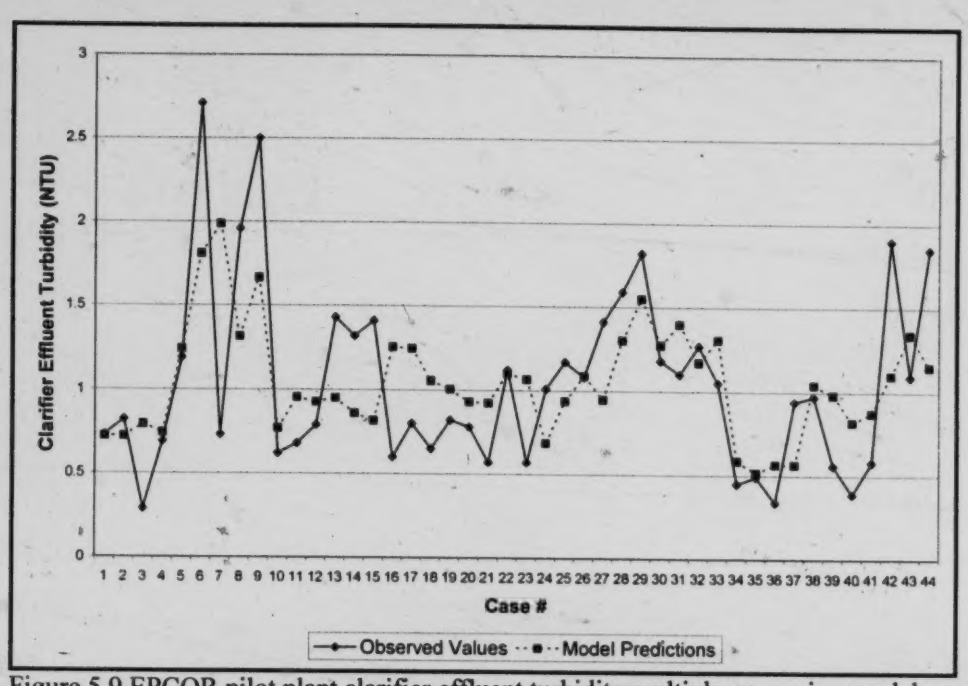


Figure 5.9 EPCOR pilot plant clarifier effluent turbidity multiple regression model results

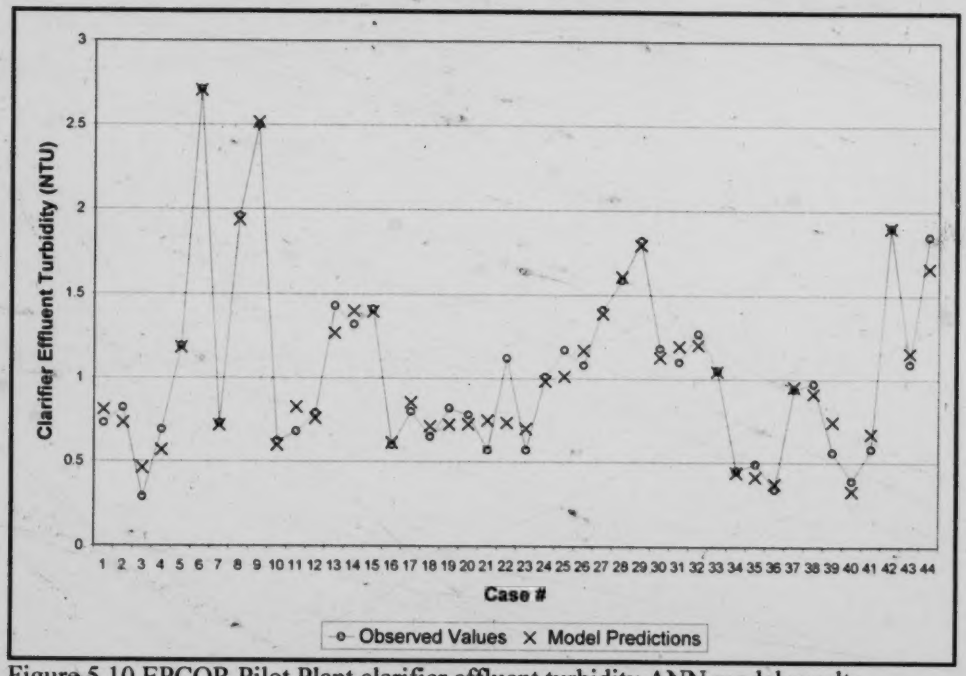


Figure 5.10 EPCOR Pilot Plant clarifier effluent turbidity ANN model results

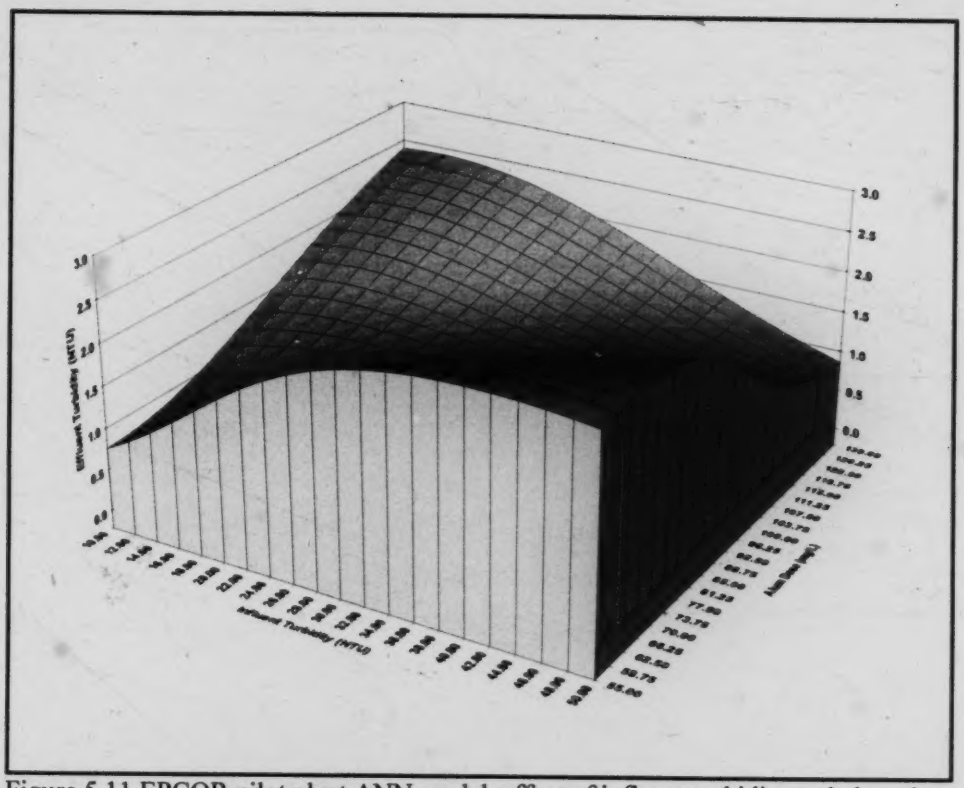


Figure 5.11 EPCOR pilot plant ANN model, effect of influent turbidity and alum dose on clarifier effluent turbidity during spring break-up conditions

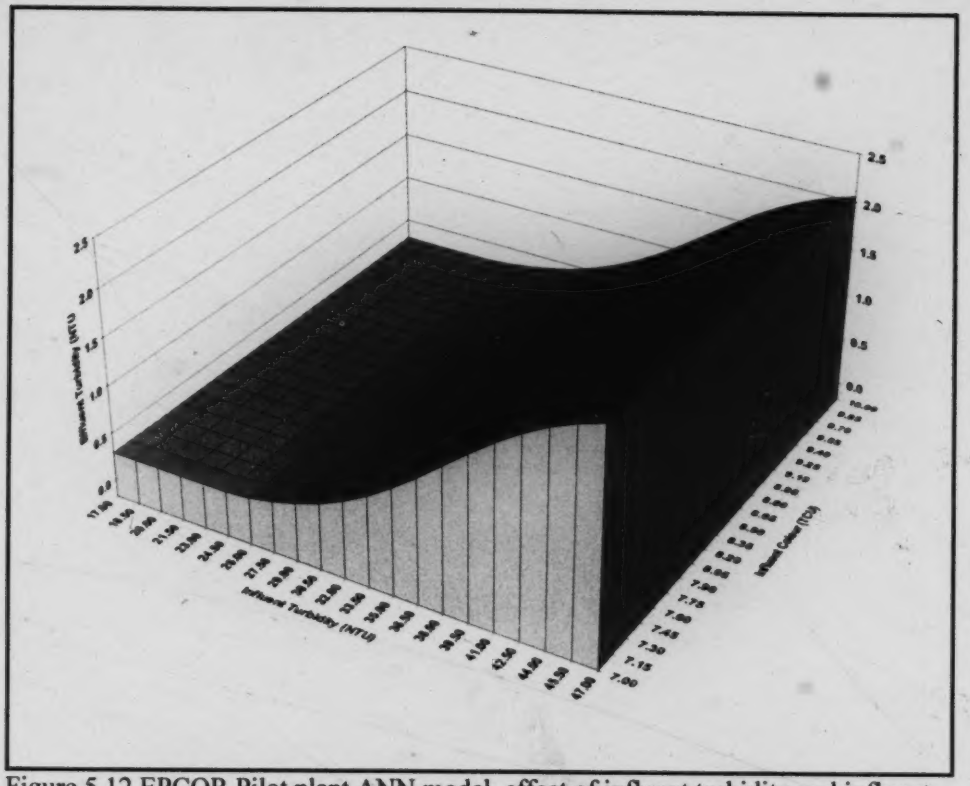


Figure 5.12 EPCOR Pilot plant ANN model, effect of influent turbidity and influent colour on clarifier effluent turbidity during typical summer operations

6. MODEL-BASED CONTROL OF ENHANCED COAGULATION*

6.1. INTRODUCTION

In the drinking water treatment industry, utilities must constantly balance profitable operations with the need to meet increasingly stringent regulatory and customer demands. This has led to many advances in plant instrumentation, as well as an increase in the use of computers in process control schemes. Currently in the industry, process instrumentation and control strategies can be divided into three levels of sophistication: supervisory control, automatic control, and advanced control (Schlenger *et al.* 1996). In supervisory control, an operator is responsible for selecting and implementing all operational changes. This control scheme is largely reactive in nature, as process changes are not made until the quality of the process effluent begins to degrade. As many unit processes have long detention times, treatment facilities are forced to endure periods of sub-optimal process performance.

Both automatic and advanced process control incorporate control logic into feed forward, feedback, or combination control schemes in order to meet operational set points. The difference between the two types of control lies in the sophistication of the control logic. In automatic control, the control logic consists of simple algorithms or linear models. In the water treatment industry, automatic control has been implemented in flow-paced

* A version of this paper has been accepted for publication. Baxter, C.W., Shariff, R., Stanley, S.J., Smith, D.W., Zhang, Q., and Saumer, E.D. Model-based advanced process control of coagulation. *Water Science and Technology*, (accepted 06/2001). 8p.

chemical feed systems and other routine control applications. The control logic in advanced control is far more sophisticated and may include complex algorithms or nonlinear process models. Advanced process control allows for proactive control decisions in real-time according to changes in any number of raw water quality or other process operating characteristics.

This chapter describes the theoretical development and practical implementation of an artificial neural network (ANN) model-based advanced process control system, believed to be the first of its kind in the water treatment industry, at a pilot-scale water treatment facility. More specifically, the design and subsequent implementation of an advanced process control system to maintain a constant clarifier effluent turbidity through the variation of alum dose, polymer dose, and plant flow in response to changes in influent water quality characteristics is described.

6.2. BACKGROUND INFORMATION

6.2.1. EPCOR Water Services Pilot Plant

The EPCOR Water Services Pilot Plant, where the study was conducted, is a research facility located on-site at the E.L. Smith Water Treatment Plant (WTP) in Edmonton, Alberta, Canada. The pilot plant draws water from the North Saskatchewan River, via the E.L. Smith facility's low-lift pump house. As a research facility, unit process operations are highly modularized; the user can specify which unit processes are included in a given

study. The plant has the ability to feed a variety of coagulants and coagulant aids, as well as powdered activated carbon, and has a maximum flow rate of up to 12 m^3/h, depending on which unit processes are included in operations. The plant is controlled via a main supervisory control and data acquisition (SCADA) computer running FIX automation software from Intellution of Foxborough, Massachusetts. The plant is outfitted with a host of online instrumentation for the analysis of raw and treated water quality, as well as the monitoring of process operating characteristics. The instruments communicate with the SCADA computer via Modicon programmable logic controllers (PLCs) and a Modbus communication network, both supplied by Schneider Electric, Inc. of Paris, France. The plant SCADA system also communicates with EPCOR's main SCADA systems via Ethernet.

6.2.2. The Coagulation Process

The primary objective of the coagulation process under study, also referred to as coagulation/flocculation/sedimentation or simply clarification, is to remove particulate matter from the treatment stream. A chemical coagulant is introduced and, through rapid mixing, is dispersed into the process water. Through slow mixing in one or more flocculation basins, the coagulant destabilizes particulate matter and promotes the formation of larger agglomerates or flocs. These larger agglomerates then settle in a sedimentation basin; this process can be enhanced through the use of tube or plate settlers. At the EPCOR Water Services pilot plant during this study, as well as at EPCOR's full-scale facilities, coagulation is achieved through the use of alum as the

primary coagulant along with an anionic polymer coagulant-aid. Sedimentation is enhanced through the use of tube settlers. The doses of alum and polymer, as well as the flow through the plant, constitute the three key manipulated variables for the process. Process performance is affected by a number of monitored raw water quality variables including the nature and concentration of particulate matter, pH, alkalinity, hardness and temperature.

6.2.3. ANN Modelling

Advances in computing power and process data handling and storage have brought a number of artificial intelligence (AI) process modelling techniques within the reach of most water treatment facilities. These techniques, which include expert systems, fuzzy logic, and artificial neural networks, are better able to handle the dynamic and non-linear nature of drinking water treatment processes. The ANN technology is the most powerful modelling tool currently available to the drinking water treatment industry. ANNs are capable of self-organization and learning; patterns and concepts can be extracted directly from historical data. When presented with data patterns, sets of historical input and output data that describe the problem to be modelled, ANNs map the cause-effect relationships between the model inputs and outputs. This mapping of input/output relationships in the ANN model architecture allows developed models to be used to predict the value of the model output variable, given any reasonable combination of model input data, with satisfactory accuracy (Baxter *et al.* 2001). Some of the more prominent past applications of the ANN technology to water treatment process modelling

include alum and polymer dose forecasting in coagulation (Mirsepassi, *et al.* 1995), and turbidity and colour removal through enhanced coagulation (Stanley *et al.* 2000).

6.2.4. Model-based Advanced Process Control

In general, advanced process control systems consist of three major integrated components; a process or process-inverse model to supply control logic, online instrumentation to provide process data, and a SCADA system computer to relay communications, execute control logic, and evaluate process performance. A conceptual framework for the use of ANNs in model-based control of the coagulation process is presented by Zhang and Stanley (1999). The proposed framework involved the use of both process and process-inverse ANN models to select an alum dose that optimizes the removal of turbidity.

6.2.4.1. Control Logic

For all but the most simple processes, advanced control cannot be effective without an empirical or mechanistic process or process-inverse model that describes the cause-effect relationships between process input and process output parameters. In the drinking water treatment industry, each unit process is governed by non-linear interactions between multiple parameters. As many of these interactions are poorly understood, no universally accepted mechanistic process models currently exist. Empirical models of drinking water

treatment processes, where they exist, are generally site-specific and are often unable to account for simultaneous changes in more than a few key process parameters.

With respect to model-based process control applications, ANNs can conceptually be incorporated in either a direct or an indirect method (Psichogios and Ungar 1991). In the indirect method, the model is a process model trained to predict the output of the process. As such, given the values of the process inputs and manipulated variables, the model predicts the expected value of the process output. In the direct method, a process-inverse model is trained to predict the value of a manipulated variable required to reach a target value of the process output. As such, given the values of the process inputs, the values of all but one of the manipulated variables, and a desired value of the process output variable, the model predicts the optimal value of a manipulated variable.

6.2.4.2. Online Analyzers

Advanced process control systems operate in real-time and are dependent upon the availability of process data in order to make effective process control adjustments. Online analyzers, which include sensors and instruments, are responsible for collecting process data on a continuous basis and relaying this information in real-time to the plant supervisory control and data acquisition (SCADA) system.

6.2.4.3. SCADA System

In an advanced process control system, the plant SCADA system is responsible for all communications between control system components, relaying control actions and process information throughout the system. The SCADA framework allows for the execution of control logic and evaluation of process performance. The system also hosts the interface between the control system and plant operators, and has the ability to archive operational data.

6.2.5. Model Integration

The application of ANN models in advanced process control requires careful attention to both hardware and software integration. Hardware integration involves the selection installation, and maintenance of online analyzers, data communication systems, and pumps and actuators (Baxter *et al.* 2001). These hardware components must be selected with the ANN technology in mind; model-based control systems require reliable real-time data in order to ensure control system integrity. Software integration of ANN models into advanced control systems can only be realized through the development of custom user and control interfaces. The control interface links with the SCADA system in real-time to retrieve model input data, execute the runtime ANN models, and transfer output data back to the SCADA system. The user interface consists of view screens and model-based applications that allow users to input control specifications and monitor system performance in real-time. A detailed discussion of interfaces and effective model

integration is presented by Baxter *et al.* (2001). The flow of information between the interfaces and the remaining components of the control system is summarized in Figure 6.1.

6.3. METHODS

6.3.1. Data Collection

The data used in ANN model development were generated through the operation of the pilot plant from November 2000 to February 2001. During this time, the levels of each of the three manipulated variables, alum dose, polymer dose, and plant flow, were randomly adjusted within pre-defined limits at set time intervals in order to generate a representative operational data set. The data for many of the raw water quality variables, as well as the process output variable used in model development, were collected online at the pilot plant and archived on the plant SCADA system. A schematic diagram of the coagulation process at the pilot plant that illustrates online data collection points is presented in Figure 6.2. As will be discussed, data for the remaining variables were obtained from either online instrument analysis or operations lab analysis at the E.L. Smith facility. The data collected on the raw water at the E.L. Smith WTP are transferable to the pilot plant as the two plants share a common raw water source, the North Saskatchewan River, and all data entries and online analyses at the E.L. Smith are time-stamped and can be synchronized with the pilot plant database. All online and

laboratory analyses were subjected to EPCOR Water Services' rigorous quality control and quality assurance protocols.

6.3.2. ANN Model Development and Control System Integration

The model development protocol presented in Chapters 2, and modified for use with Statistica Neural Networks as described in Chapter 3, was employed. Data analysis was accomplished using Microsoft Excel 2000, while ANN model development was carried out using StatSoft's Statistica Neural Networks. The control logic provided by the ANN models was interfaced with the EPCOR Water Service's pilot plant SCADA system using custom-designed Microsoft Excel spreadsheet control and user interfaces, as will be discussed.

6.4. RESULTS AND DISCUSSION

6.4.1. ANN Model Development and Evaluation

Successful ANN models can only be developed if representative data for the process to be modelled exist. Based on the author's experience in modelling the coagulation process at full-scale WTPs, described by Stanley *et al.* (2000), a number of data variables were identified for inclusion in the pilot plant model. As the pilot plant is a research facility, the quality and quantity of data collected during each study are typically researcher specific. A survey of coagulation studies previously conducted at the pilot plant failed to

yield a useable data set for ANN modelling. As such, a data collection program was developed and implemented on November 27th 2000 and data collection continued through February 26th 2001. A list of the variables collected for use in model development, along with the source, mean, and standard deviation of each variable, is presented in Table 6.1. In order to collect as much relevant data as possible, the process operating characteristics were altered every four or six hours, depending on plant flow. The doses of alum and polymer were selected using a random dose selection algorithm, programmed by a member of the EPCOR Water Services Controls Group, that varied the polymer dose between 0.05 and 0.55 mg/L and the alum dose between 5 and 55 mg/L. These limits were selected to correspond with typical dosing ranges at EPCOR's full–scale facilities during the winter months when the North Saskatchewan River is under ice cover. During data collection, the plant flow was also varied periodically between 1.75 and 4.25 m^3/h. Initially, these changes were implemented manually, however, at the end of December, online plant flow control was implemented and flow selection was incorporated into the random dose selection algorithm.

Over the course of the study, three separate ANN models were sequentially used to supply the control logic for the advanced process control system. As will be discussed, an improved model was developed whenever model deficiencies were identified. As such, data generated during the first month of the system's operation was used in subsequent model development. Each model had a three-layer multi-layer perceptron architecture with 9 hidden layer neurons and was trained using the backpropagation and conjugate gradient descent learning rules described in Chapters 2 and 3. The results of each model,

when applied to training, testing, and independent validation sets are presented in Table 6.2. The independent validation data sets consist of data not involved in model training and serve to ensure that the models do not simply memorize the data used in training and internal model testing, as discussed by Stanley *et al.* (2000). A sample modelling data set, which demonstrates the distribution of data among the three data sets, is presented in Appendix A (Table A.1).

6.4.2. Control System Integration

The trained ANN models were incorporated into a model-based advanced process control system using an indirect, or feed forward, methodology. The system varies the levels of manipulated variables (alum dose, polymer dose, and plant flow) in response to changes in process variables (raw water quality variables) in order to achieve and maintain a target value of the controlled variable (clarifier effluent turbidity). As previously discussed, the advanced process control system consisted of three major integrated components. In order for the system to be successful, it was necessary to ensure effective integration. For the purposes of the current study, integration was accomplished using a Microsoft Excel spreadsheet user interface coupled with a custom Microsoft Visual Basic control interface. Over the course of the study, two sets of interfaces were developed and deployed. The original interfaces allowed the user to specify a target value for clarifier effluent turbidity as well as high and low limits for alum dose and polymer dose. The improved interface introduced additional functionality: user defined low, high, and target values for plant flow, process optimization via prioritization of flow, profit, or clarifier

effluent turbidity targets, and user defined tolerances for all manipulated and controlled variables.

While a detailed description of the intricacies of the interfaces and related calculations is beyond the scope of the present discussion, a general discussion of interface functionality is warranted, and is given here in reference to the improved interface. The Microsoft Visual Basic control interface, programmed by a member of the EPCOR Water Services Controls Group, deploys an algorithm to execute the control logic. A simplified schematic of the algorithm is presented in Figure 6.3 and is discussed here. The control interface initially acquires current values for each of the raw water quality model inputs from the plant SCADA system, as well as user specifications regarding the prioritization, target values, and tolerances for each of the manipulated and controlled variables. To simplify the current discussion, the algorithm will be discussed in reference to the situation where the user wishes to prioritize operations first according to effluent turbidity, next by plant flow, and finally by profit. The interface superimposes the raw water quality information on a three-dimensional array that defines all conceivable combinations of alum dose, polymer dose, and plant flow. While the user can define how many combinations are evaluated, a 50 by 50 by 50 array (125000 combinations) equally distributed across the ranges of the three manipulated variables provides ample resolution for the control system. The ANN process model is run 125000 times and generates predictions using the raw water quality information and the 125000 combinations of alum dose, polymer dose, and plant flow. At the first decision point, all runs that result in a clarifier effluent turbidity that falls between the upper and lower user specified limits are

identified and passed on to the second decision point. If the clarifier effluent turbidity target is not met by any of the runs, the current operational conditions are maintained and the algorithm stops. At the second decision point, all the remaining runs that fall between the user-defined lower and upper limits for plant flow are identified and passed on. If the flow target cannot be met, the operational conditions that will come closest to meeting the flow target are applied and the algorithm stops. In the final process, operations are optimized by selecting the most cost-efficient combination of alum dose and polymer dose that will allow both the clarifier effluent turbidity and plant flow targets to be met. The values of alum dose, polymer dose, and plant flow that result from this operation are returned to the plant SCADA system for immediate implementation. Due to advances in computing power, the processing time for the control algorithm is less than 5 seconds on a personal computer with a 600 MHz Pentium III processor with 128 M of RAM.

6.4.3. Control System Evaluation

In order to demonstrate the utility of the advanced process control system, a number of different case studies are discussed. The first case involves the original interface, uses Model 1 to supply the control logic, and involves plant operations from February 7th to February 15th 2001. The clarifier effluent turbidity target was set to 0.25, which is lower than possible under all but the most favorable raw water quality conditions. The control system therefore attempts to minimize clarifier effluent turbidity. As can be seen in Figure 6.4, the control of clarifier effluent turbidity is extremely stable in spite of large fluctuations in influent particle counts. As part of the data analysis phase of the study,

clarifier effluent turbidity was found to be highly correlated to raw water particle counts ($r = 0.70$) when no attempt to optimize the process for particulate removal was made. At 10:30 a.m. on February 9th, the flow through the plant was manually reduced to 2.25 m^3/h from 2.50 m^3/h. As was previously discussed, the original interface did not include provisions for automated selection of optimal plant flow. Nevertheless, the control system easily accommodated the new lower flow and the resulting increase in detention time allowed for better particle removal. The small gap in clarifier effluent turbidity measurements during the transition from one flow to another can be traced to a power spike at the pilot plant that tripped the pump that supplies the on-line turbidity meter. Data collection resumed once the pump was brought back on-line.

Encouraged by the preliminary results obtained using the original interface, a more sophisticated control system interface was developed and deployed for the remaining four case studies. In order to simplify the discussion of these case studies, the conditions and results of the studies have been grouped together and are presented in Table 6.3. As was previously discussed, the improved interface allows the user to define and prioritize targets for clarifier effluent turbidity, flow, and profit. The last is determined using a typical EPCOR residential water rate ($\$0.942/m^3$) along with unit costs of alum and polymer. For each of the four case studies, the order of prioritization for the user defined targets, from most to least important, was: clarifier effluent turbidity, plant flow, and profit. As such, in determining the optimal levels of the three manipulated variables, the control system first identifies all the combinations from the three-dimensional array that satisfy the clarifier effluent turbidity target. Of these combinations, those that meet the

flow criteria are identified and, finally, the combination that is the most cost efficient is selected for implementation. If an optimal solution can't be identified, the system defaults to the last known optimal solution, as previously discussed.

In Case 2, which represents operations from February 23rd to 26th, the clarifier effluent turbidity and plant flow targets were set to 2.50 NTU and 3.00 m^3/h, respectively. While the plant flow target was met throughout the study, the ANN model consistently over-predicted the value of clarifier effluent turbidity. As a result, the observed values of clarifier effluent turbidity are negatively offset from the target values. In order to correct the offset, a new model (Model 2) was trained with additional data collected during the first case study.

In Case 3, which uses Model 2 for control logic and covers operations from February 28th to March 2nd, the clarifier effluent turbidity and plant flow targets were set to 2.00 NTU and 3.50 m^3/h, respectively. Using the new model, the control system results improved dramatically. A mean clarifier effluent turbidity of 2.03 NTU with a standard deviation of 0.20 NTU was obtained over the course of the test. As can be seen in Figure 6.5, the flow target was only met when the raw water particle counts were below approximately 13,000 counts/mL. A further examination of the test data revealed that control was maintained solely through polymer dose and plant flow variation in response to changes in raw water quality. In EPCOR Water Service's full-scale facilities during the weeks preceding spring thaw, increases in alum dose have historically resulted in increased clarifier effluent turbidity as the particulate matter is not easily destabilized by the metal coagulant in cold

high-alkalinity water. This operational feature was captured during the development of Model 2, which predicts increases in clarifier effluent turbidity with alum dose. As a result, alum doses were minimized by the system throughout the test. The system recognized that the only way to maintain clarifier effluent turbidity control was to decrease plant flow during high particle count events.

Immediately following Case 3, the ambient temperature in the Edmonton area became unseasonably warm, resulting in the partial melting of the North Saskatchewan River's ice cover. Early spring thaw conditions are typically accompanied by a rapid deterioration in raw water quality at the EPCOR treatment facilities as organic matter is introduced into the river through snowmelt. This period of transition is also characterized by an increase in the effectiveness of alum in the removal of particulate matter as lower alkalinity values result in improved coagulation pH conditions, and water temperatures increase. In order to ensure continued successful operation of the advanced process control system, a new model (Model 3) that predicts a decrease in clarifier effluent turbidity with an increase in alum dose was developed. The new model was incorporated at the start of Case 4, which represents plant operations from March 12th to 19th. The values of clarifier effluent turbidity and plant flow for the test were targeted at 2.00 NTU and 3.00 m^3/h, respectively. An in-depth analysis of the process data and subsequent testing revealed that the new model was implemented prematurely. Each day during the test, raw water particle counts, clarifier effluent turbidity, and alum dose demonstrated a cyclical behaviour, peaking at approximately 11:00 p.m. and reaching their lowest values just after 6:00 a.m. The particle count cycle was found to correlate with the daily freeze-

thaw cycle that is prevalent during spring thaw. The clarifier effluent turbidity cycle was found to be highly correlated with both the particle counts cycle and the alum dose cycle. In order to determine whether the clarifier effluent peaks were being caused by increases in alum dose, the alum dose was restricted to a value of 6 mg/L on two successive days. During this time frame, the daily particle count cycles continued, however, the clarifier effluent turbidity did not follow suit.

The final case study, representing plant operations between March 23rd and 25th again involved the use of Model 3 in an attempt to reach a stringent target clarifier effluent turbidity of 1.75 and a target plant flow of 2.75 m^3/h. Sometime between the end of Case 4 and the start of Case 5, the anticipated change in raw water particulate characteristics occurred. Large agglomerates, characteristic of effective alum coagulation, were observed in the flocculation basins and tube settlers of the clarifier. As can be observed in Figure 6.6, the clarifier effluent turbidity shows little variation over the course of the case study. The discrepancy at the start of the test is a result of bringing the control system back online after four days of inactivation. Control was accomplished by varying alum dose and polymer dose in response to the cyclical raw water quality changes. The control system tended to over-predict clarifier effluent turbidity, resulting in observed clarifier effluent turbidity values that were negatively offset from the target. This offset can be corrected through further model development as more operational data become available. The plant flow target was met with a mean observed flow of 2.78 m^3/h and a standard deviation of 0.04 m^3/h.

6.5. CONCLUSIONS

The results of this pilot project highlight the tremendous potential for the development and application of ANN model-based advanced process control systems in the drinking water treatment industry. The encouraging results generated from the first case study where tight control of clarifier effluent turbidity was achieved led to the development of a sophisticated interface for the advanced process control system. Using ANN models as the control logic, this system allowed the user to define and prioritize target values for key process operating characteristics.

Prior to full-scale implementation of the technology, further investigation of system robustness under a variety of raw water quality conditions, as well as the development of systems for other unit processes, is required. The technology will also benefit from a thorough investigation of security protocols associated with model-based process operations, as well as quality assurance issues surrounding process data collection and communication between system components. The information gained through such studies will serve to highlight potential challenges and concerns associated with full-scale implementation, and will generate increased confidence in the use of the technology in the drinking water treatment industry.

6.6. REFERENCES

Baxter, C.W., Zhang, Q., Stanley, S.J., Shariff, R., Tupas, R-R. T., and Stark, H.L. (2001). Drinking water quality and treatment: the use of artificial neural networks. *Can. J. Civ. Eng.* 28(Suppl. 1): 26-35.

Mirsepassi, A., Cathers, B., and Dharmappa, H.B. (1995). Application of artificial Neural networks to the real time operation of water treatment plants. In *IEEE International Conference on Neural Networks: Proceedings*. Perth, Australia: IEEE

Psichogios, D. C., and Ungar, L. H. (1991). Direct and indirect model based control using artificial neural networks. *Ind. Eng. Chem. Res.* 30: 2564-2573.

Schlenger, D.L., Riddle, W.F., Luck, B.K., and Winter, M.H. (1996). *Automation Management Strategies for Water Treatment Facilities*. Denver, CO.:AWWARF and AWWA. 195 p.

Stanley, S.J., Baxter, C.W., Zhang, Q., and Shariff, R. 2000. *Process Modelling and Control of Enhanced Coagulation*. American Water Works Association Research Foundation and American Water Works Association, Denver, CO: 167 p.

Zhang, Q. and Stanley, S.J. (1999). Real-time water treatment process control with artificial neural networks. *J. Env. Eng.* 125(2):152-160.

Table 6.1 ANN model variables for the advanced process control system

Variable	Source	Mean	Std. Dev.
Temperature (°C)	Online, pilot plant	1.64	0.14
Particle counts (counts > 2 μm/mL)	Online, E.L. Smith	9776.50	3119.36
Colour (TCU)	Lab, E.L. Smith	4.49	1.04
Alkalinity (mg/L)	Lab, E.L. Smith	135.03	9.28
pH	Online, E.L. Smith	8.08	0.11
Hardness (mg/L as $CaCO_3$)	Lab, E.L. Smith	179.98	10.71
Plant flow (m^3/h)	Online, pilot plant	2.86	0.63
Alum dose (mg/L)	Online, pilot plant	30.28	17.84
Polymer dose (mg/L)	Online, pilot plant	0.30	0.17
Clarifier effluent turbidity (NTU)	Online, pilot plant	1.66	1.26

Table 6.2 ANN model results for the advanced process control system

Model	Training data		Testing data		Validation data	
	R^2	MAE (NTU)	R^2	MAE (NTU)	R^2	MAE (NTU)
1	0.94	0.23	0.94	0.26	0.92	0.27
2	0.94	0.25	0.94	0.23	0.85	0.34
3	0.90	0.28	0.94	0.26	0.81	0.36

Table 6.3 Target values and results for the control system; Case Studies 2 to 5

Case	Model	Target values		Turbidity (NTU)		Flow (m^3/h)		Profit ($/$m^3$)
		Turbidity (NTU)	Flow (m^3/h)	Mean	Std. Dev.	Mean	Std. Dev.	Mean
2	1	2.50	3.00	2.07	0.35	3.04	0.20	0.936
3	2	2.00	3.50	2.03	0.20	2.61	0.83	0.939
4	3	2.00	3.00	2.26	0.43	2.99	0.09	0.937
5	3	1.75	2.75	1.36	0.22	2.78	0.04	0.939

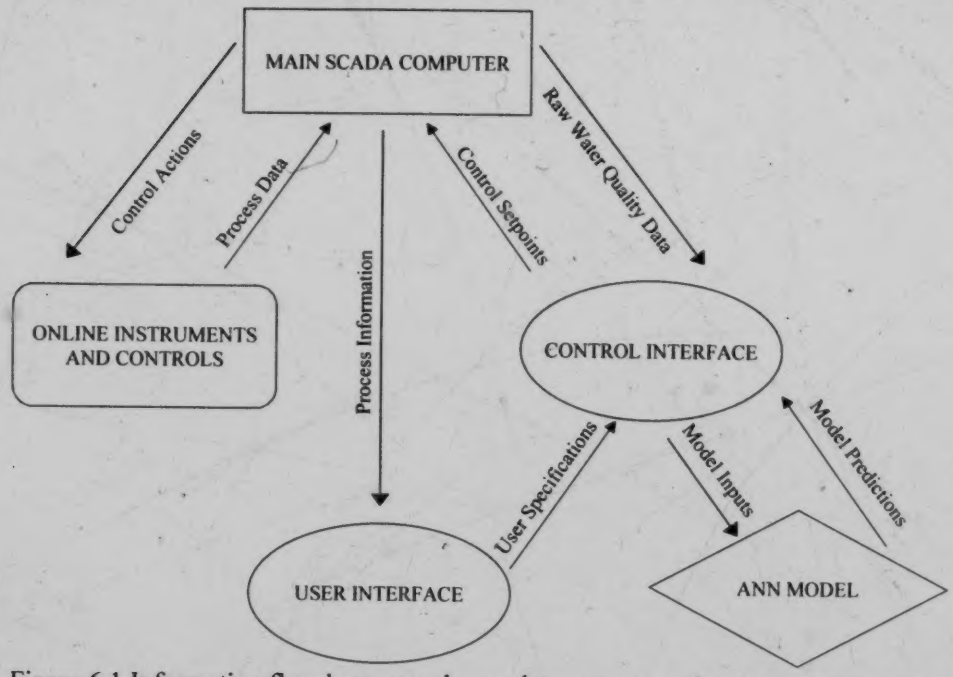


Figure 6.1 Information flow between advanced process control system components

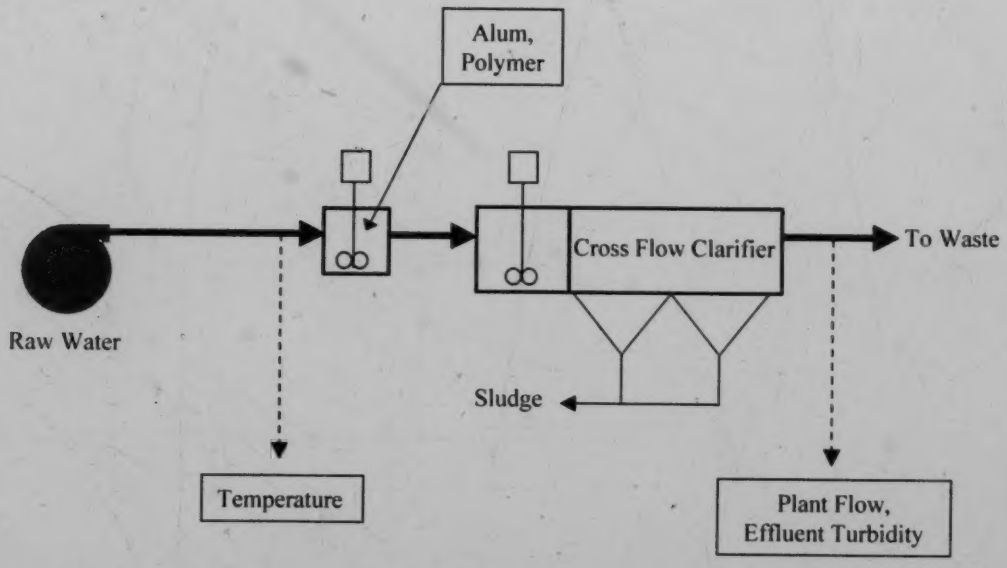


Figure 6.2 Schematic diagram of the EPCOR pilot plant coagulation process

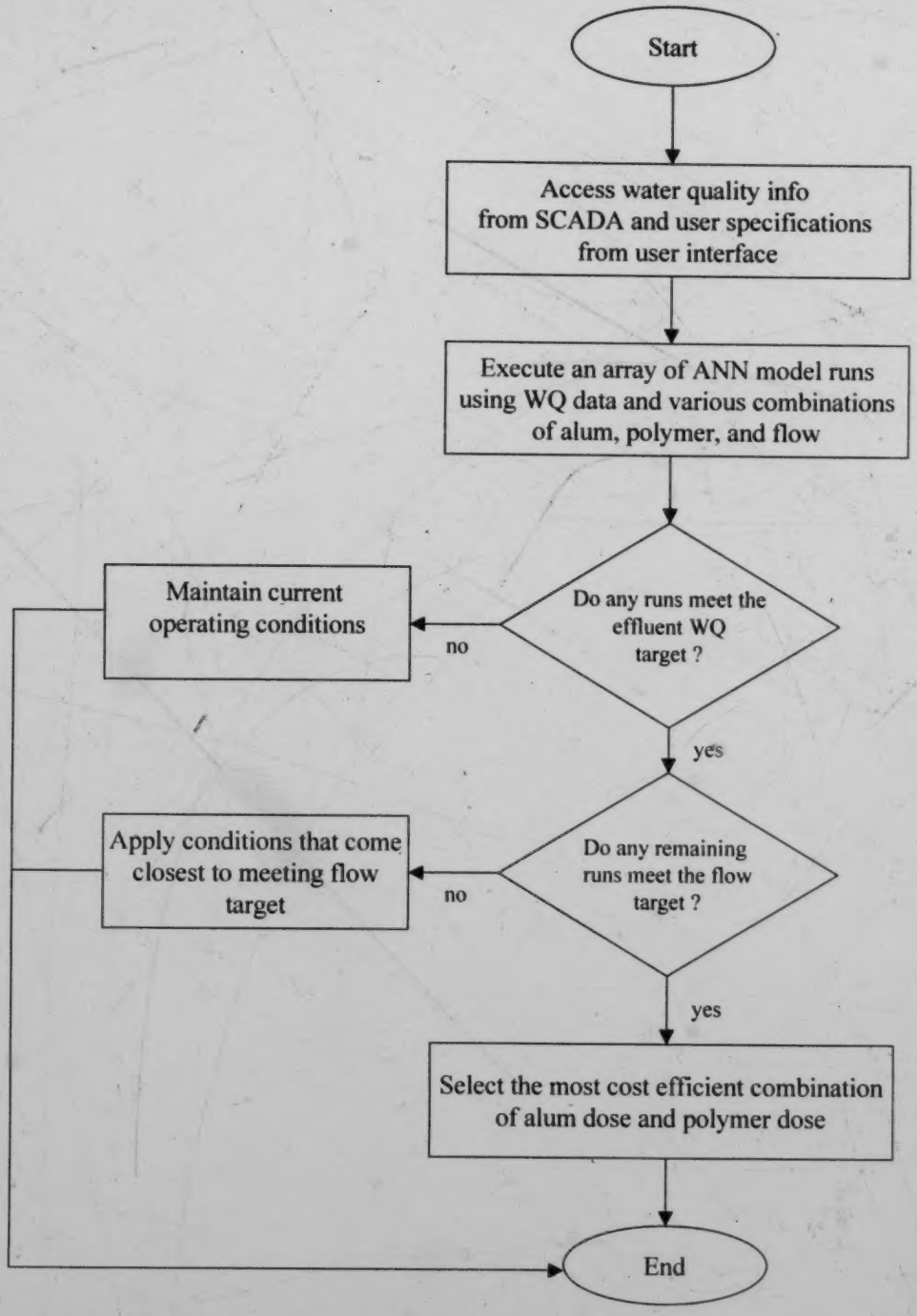

Start
Access water quality info from SCADA and user specifications from user interface
Execute an array of ANN model runs using WQ data and various combinations of alum, polymer, and flow
Do any runs meet the effluent WQ target ?
no
Maintain current operating conditions
yes
Do any remaining runs meet the flow target ?
no
Apply conditions that come closest to meeting flow target
yes
Select the most cost efficient combination of alum dose and polymer dose
End

Figure 6.3 Simplified version of the control system algorithm

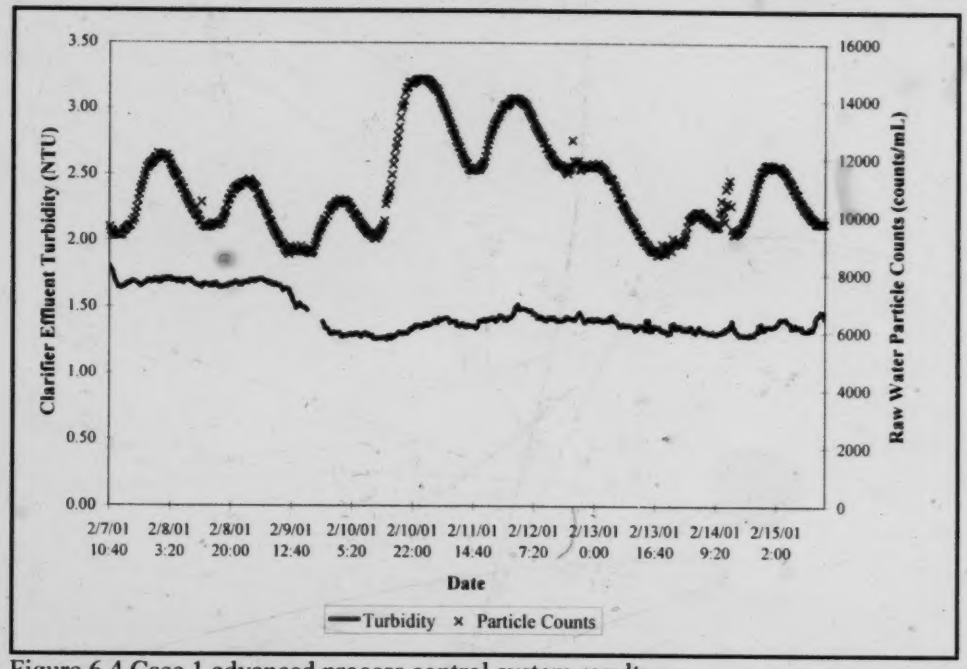


Figure 6.4 Case 1 advanced process control system results

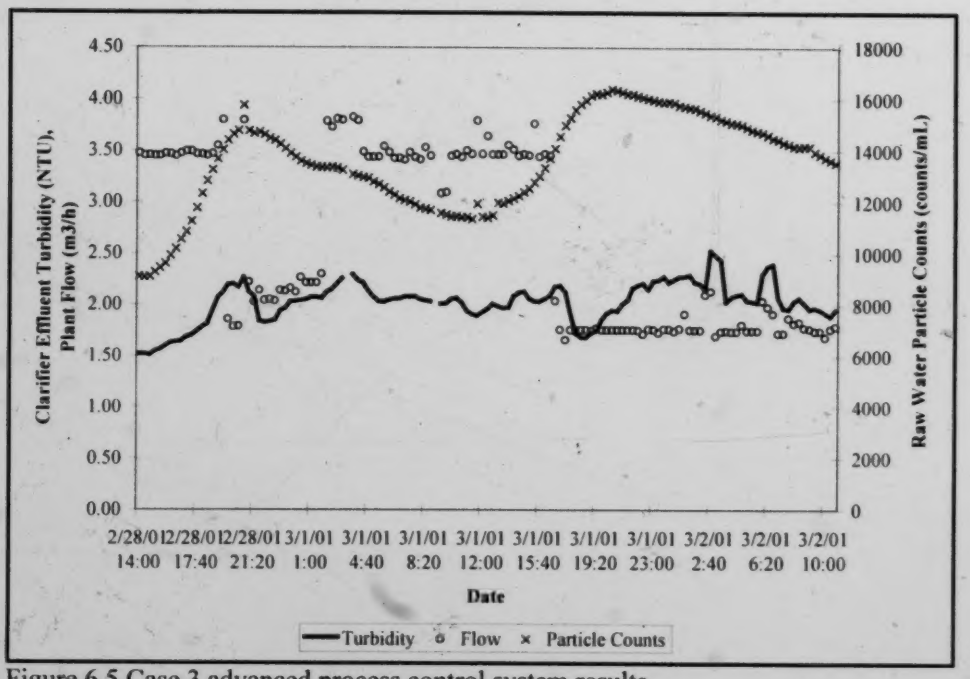


Figure 6.5 Case 3 advanced process control system results

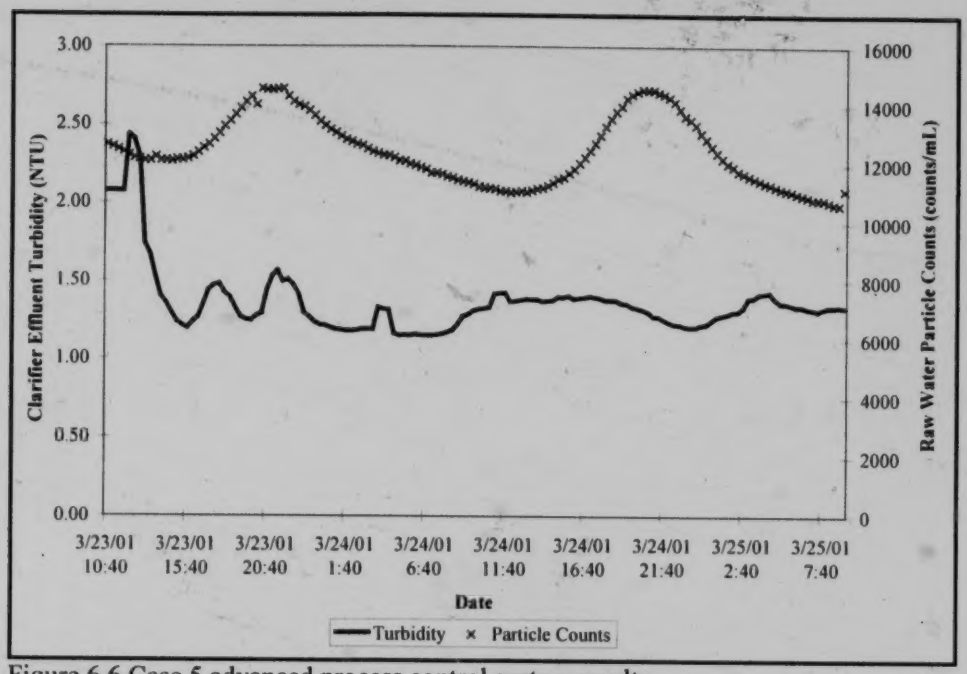


Figure 6.6 Case 5 advanced process control system results

7. GENERAL DISCUSSION AND CONCLUSISONS*

7.1. INTRODUCTION

The overall goal of the research program was to apply the ANN technology to process modelling and control in the drinking water treatment industry and assess model performance in the developed applications. Process models were successfully developed for a wide variety of treatment facilities with equally variable raw water and operational characteristics. Both online and offline operational tools using developed models were summarily discussed. In addition, new advances in protocols for model development, the evaluation of model boundaries, the application of ANNs in pilot-scale data analysis, and ANN-based advanced process control were presented.

In this concluding chapter, the applicability of each developed ANN technology and application to water treatment operations is summarily discussed with a focus on the practical aspects of implementation. In addition, a summary of the major research findings from the various chapters is presented. Finally, recommendations for further study in reference to potential future directions of water treatment process control are made.

---

* A version of this chapter has been published. Baxter, C.W., Tupas, R-R.T., Zhang, Q., Shariff, R., Stanley, S.J., Coffey, B.M., and Graff, K.G. 2001. *Artificial Intelligence Systems for Water Treatment Plant Optimization.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 141 p.

7.2. DISCUSSION AND APPLICATION POTENTIAL

A common misconception about the ANN technique is that it is only applicable to large treatment facilities with reams of historical data and requires extensive financial commitment on the part of utilities. All treatment facilities that have some historical data can apply the ANN modelling technique. As will be discussed, increasingly sophisticated models and applications can be developed if certain process control and SCADA system requirements are met.

7.2.1. Process Assessment and Data Analysis

The least complex application of the ANN technique in the drinking water treatment industry involves the use of ANN models for process assessment and data analysis. While models can be developed for virtually any amount of data, the success of models in mapping key relationships in treatment processes can only be ensured when a representative data set is used. As was demonstrated in Chapter 3, successful models were developed using data sets that contained fewer than 80 patterns. In the ODP Plant models described in Chapter 3, the data sets were carefully evaluated to ensure that they were indeed representative of the conditions under which the models were to be applied.

With regards to the practical implementation of the ANN technology for process assessment and data analysis applications, any treatment process can be modelled providing that sufficient representative data exist. In practice, for processes that exhibit

seasonal variation in water quality and performance, data that spans at least one full-year of operations should be considered a minimum requirement. In order to identify parameters to be used in modelling a given process, a review of current literature on the process is required. An assessment of the utility's historical data records can then be made and can be referenced to the results of the literature review. The utility can then determine whether sufficient data exist for the development of successful models and applications.

Other than data requirements, all that is required to develop simple models for process assessment and data analysis is ANN modelling software and a user that is knowledgeable in both water treatment operations and ANN modelling. With respect to the former, two software packages were used to develop all models presented in this report: NeuroShell 2 from Ward Systems Group, Inc. and Statistica Neural Networks, from StatSoft, Inc. These programs combine user friendliness with a high degree of user input. Both can also be used to generate runtime versions of the models for use in offline and online applications. With respect to the latter, university researchers, research engineers, and process engineers are all suitable candidates for developing and implementing ANN models in the water treatment industry. Many of the ANN software manufacturers offer training courses for their products and ensure that the user is making use of the full potential of the ANN modelling technique. Depending on the availability of in-house expertise, process assessment tools can be developed without additional formal training.

The cost of developing process assessment and data analysis applications is dependent on the degree of in-house modelling expertise, the availability of computing resources, and the selection of modelling software. Models can typically be developed using a single standard desktop or laptop computer, as previously discussed, and modelling software can typically be acquired for less than $ 2000. The time required to develop successful applications is dependent on the availability and format of modelling data, as well as the modeller's knowledge of the process being modelled and the ANN modelling technology. If a utility has a well-organized and easily accessible historical database and has taken the time to train a process engineer on the use of the ANN technology, it is not unreasonable to expect that successful models could be developed within a three-month time frame. As each modelling project is confronted with unique data challenges, developing process assessment applications is site and process specific. As such, the time estimate is subject to a high degree of variation and should not be over-emphasized.

As presented in this work, ANN models can be used to identify and assess difficulties in plant operations and suggest potential remedies. The technique can also assist operators in determining the effects of typical operating conditions on a newly measured treated water parameter. As utilities switch from relying solely on turbidity for filter performance assessment to newer measurement technologies such as particle counts, for example, operational strategies are subject to change. ANN models can serve to highlight the differences in operations required to optimize particle count removals, as compared to turbidity removals.

The application of the ANN technique to the analysis of pilot plant data, as presented in Chapter 5, is another application of limited complexity. This application is well suited to utilities that wish to become more comfortable with the technology, and those that already use the technology, alike. As demonstrated, the technique was able to extract useful information about the effect of a variety of parameters on process performance from extremely limited data sets. The technique takes advantage of well-designed pilot-scale experiments and observational data to map meaningful relationships between process inputs and outputs. Conventional statistical methods of data analysis, such as multiple regression modelling, have much greater difficulty separating the effects of the controlled and random variables and therefore provide less useful information concerning process operations.

7.2.2. Offline Operational Tools

Utilities that have a more extensive historical database consisting of a wide variety of reliable water quality and operational data can use the ANN technique to develop a number of offline tools to assist operators in daily plant operations. Representative tools discussed in this work include scenario analysis and virtual laboratory applications. Such applications use the historical data to develop more widely applicable relationships than possible with process assessment models. An example of a possible application of this type that has been presented in this work involved the use of the virtual laboratory for the analysis of the effect of alum and polymer on filter effluent particle counts as described in Chapter 3. Scenario analysis applications are also possible and can help answer

questions concerning the cause of process upsets. By identifying the variables responsible for a historical process upset and subsequently analyzing the effect of these variables over a given range, alternative operational changes for future scenarios can be avoided.

7.2.3. Online Operational Tools

The same types of applications identified in the previous section can be integrated into a SCADA system for execution with real-time data. In addition to the requirements previously identified, utilities wishing to pursue this level of application must have a SCADA system that is capable of interfacing with historical data and the runtime ANN models. More specifically, data for each of the model parameters must be available to the model in real-time. This generally, although not always, implies the use of online analyzers for data acquisition. Using integrated real-time online ANN models, plant operators can take advantage of a host of tools to assist them in daily plant operations. The utility of the online tools can be hampered by ineffective integration; the considerations briefly outlined in Chapter 6 must be implemented in order to make the best use of the ANN technology. Representative operations include real-time scenario analysis whereby operators determine the effect of proposed control actions on the current state of the system. Another popular online operational tool involves the use of trained ANN models in raw water quality or water demand forecasting. Such applications can assist operators in anticipating variations in raw water quality while ensuring an adequate supply of finished water.

7.2.4. Advanced Process Control

The most sophisticated level of ANN applications involves the use of trained models in real-time advanced process control. As discussed in Chapter 6, a trained ANN model can be successfully used as the control logic in a feed-forward model-based control system. The difference between automatic control and advanced control lies in the sophistication of the control logic. A flow-paced chemical feed system, for example, constitutes automatic control. On the other hand, using ANN models to automatically select the most efficient combination of alum and polymer doses in clarification in response to changes in raw water quality and operational characteristics constitutes advanced process control. The requirements for this type of application include a reliable SCADA system-based control system framework and reliable data. More specifically, contingencies for online analyzer quality assurance and quality control, error detection and alarming, system failures and back-up, and an efficient communications network must be addressed. Utilities willing to pursue this type of system must also ensure that operations staff are sufficiently trained in the use of the ANN technology as well as control systems maintenance and operations. As current regulatory and water quality pressures drive the drinking water industry to more sophisticated levels of process control, ANN-based advanced process control has the potential to become a strategy of great importance.

7.3. SUMMARY OF MAJOR FINDINGS

The ANN technique is quickly moving beyond limited use as a research tool to widespread applicability in the optimization of full-scale water treatment operations. Through numerous model applications, this work reaffirms that the technique is indeed feasible and is therefore recommended for continued use in the industry. Specific conclusions based on the results of the work include:

1. ANN models can best be developed using a systematic methodology. The steps involved in model development include a needs and suitability analysis, data collection and analysis, model development protocol application, and model evaluation.
2. The ANN technique can be used for process assessment using small data sets as long as the data are representative of the conditions on which the trained model is to be applied
3. The technique is equally applicable to large representative data sets. Difficulties can arise, however, when an attempt is made to model a process with little variation as evidenced by the results of the F.E. Weymouth Filtration Plant filter effluent turbidity model.
4. Model applications for process optimization include both online and offline operational tools. These tools serve to assist operators in plant operations through highlighting and determining the effects of key process variables.

Such tools are also extremely useful in scenario analysis and in virtual laboratory applications.

5. The applicability of ANN models to data outside the training domain is affected both by the data ranges of key model variables and model training characteristics. While quantitative relationships between the distance of a data value to the training domain and prediction results can be established, such relationships are believed to be site specific.
6. The ANN technique is particularly useful at extracting meaningful relationships from limited pilot-scale data sets. ANN models can incorporate changes in fixed and random variables simultaneously, an advantage not matched by traditional methods of statistical data analysis.
7. ANN models can successfully serve as the control logic in model-based advanced process control systems. Advanced process control systems require careful attention to model integration and development of user and control interfaces.

7.4. RECOMMENDATIONS FOR FUTURE STUDY

While this work presents a number of new developments concerning the use of ANN models in water treatment process modelling and control, several areas of exploration were beyond the scope of study. More specifically, in order to encourage and further the growth of this technology in the drinking water treatment industry, it is recommended that the following issues be resolved through future study:

1. Kohonen ANNs, briefly discussed in Chapters 2 and 6, have the potential to find increased use in the classification of model input data, as well as in data error detection. Further work on KNN protocol development, as well as the uses and limitations of the technique, is required.
2. As was evidenced by the results presented in Chapter 5, each ANN software package introduces its own unique features and limitations to model development. Since most of the work surrounding ANNs in the water treatment industry has centered around the use of NeuroShell 2 and Statistica Neural Networks, the features and limitations of other software packages remain to be addressed. It is therefore recommended that a comprehensive examination of other ANN software packages be performed so that existing model development protocols can be appropriately modified.
3. Since real-time process control applications are currently in limited use in the industry, an assessment of the robustness of current interfaces and applications over a longer time frame is recommended. This assessment would serve to highlight deficiencies in existing interfaces and recommend a prudent course of action for ensuring the long-term viability of ANN applications.
4. The advanced process control system presented in Chapter 6, believed to be the first of its kind in the water treatment industry, was developed for a single pilot-scale facility. As each treatment facility will undoubtedly present its own unique control challenges, the development of additional control systems is

warranted. In particular, advanced process control systems should be developed for facilities that use other types of source water and different treatment processes that those presented herein. The technology, while promising, will only be fully validated following a series of full-scale implementation studies.

5. Looking further into the future, a promising direction for water treatment process control will be the development of plant-wide control systems where water quality and the economics of treatment are managed from source to tap. In order for ANNs to factor prominently in plant-wide control, protocols for linking and integrating multiple models will need to be developed

7.5. CONCLUSION

The future of the drinking water treatment industry will be characterized by a more stringent regulatory environment as well as increased pressures to provide efficient treatment. In response to these changes, new technologies are continually being developed, evaluated, and implemented in the industry. The ANN technology has been applied sporadically in drinking water quality and treatment applications over the last decade and, while applications have generally been successful, the technology has yet to see widespread use in the industry.

The results presented in this work will hopefully alleviate some of the concerns that are currently causing barriers to widespread implementation of the ANN technology. A

simple and systematic protocol for developing ANN process models was developed and presented. As well, models of filter effluent and clarifier effluent performance at pilot-scale and full-scale facilities in Canada and the United States of America were successfully developed. New insights into modelling limitations were gained from these modelling efforts, and successful online and offline process optimization applications were evaluated. Model prediction boundaries were quantified, and the importance of various model training parameters on model prediction ability outside the training domain were investigated. Finally, new applications of the technology in the analysis of pilot-scale data and in real-time advanced process control were developed, implemented, and evaluated.

In combination, the results presented herein have the potential to become the definitive body of knowledge for developing ANN applications in water treatment process modelling and control. Future ANN technologies in the industry can use the work as a springboard for more sophisticated applications than are currently possible. With further study, the ANN technology will undoubtedly play a substantial role in assisting utilities to meet future regulatory and economic demands.

APPENDIX A – MODELLING DATA

Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3

Date and Time	Set	Inf. Temp. (deg. C)	Inf. P.C. (counts/mL)	Inf. Colour (TCU)	Inf. Alk. (mg/L)	Inf. pH	Inf. Hard. (mg/L)	Flow (m3/h)	Alum (mg/L)	Polymer (mg/L)	Eff. Turb. (NTU)
11/27/00 16:00	Train	1.54	8046	4.40	140	8.22	178	3.95	20	0.35	0.86
11/27/00 19:00	Train	1.52	7717	4.06	140	8.20	178	3.95	30	0.35	0.65
11/27/00 22:40	Prod.	1.50	7216	4.06	138	8.18	178	3.95	30	0.35	0.96
11/28/00 9:20	Train	1.49	7155	3.40	136	8.20	178	3.89	30	0.35	0.82
11/28/00 10:00	Train	1.49	7200	3.40	136	8.04	178	2.05	40	0.35	0.80
11/28/00 15:00	Train	2.05	8413	3.40	136	8.36	178	3.93	10	0.25	0.97
11/29/00 8:40	Train	1.70	9768	3.70	136	8.21	178	3.89	9	0.25	0.79
11/30/00 11:00	Train	1.52	9609	2.60	136	8.24	180	2.49	30	0.10	1.05
11/30/00 13:40	Train	1.56	8571	2.60	136	8.22	182	3.06	50	0.50	0.97
11/30/00 15:40	Train	1.54	9194	2.60	136	8.23	182	3.06	40	0.50	0.77
11/30/00 16:20	Prod.	1.54	8987	2.60	136	8.23	182	3.06	40	0.10	0.76
11/30/00 19:40	Train	1.53	8291	2.60	136	8.23	182	3.01	40	0.10	1.30
12/1/00 7:40	Test	1.54	7900	5.70	138	8.25	182	3.01	40	0.10	1.29
12/1/00 9:40	Train	1.53	7643	5.70	138	8.25	182	3.01	10	0.35	1.00
12/1/00 17:40	Train	1.48	10208	4.96	138	8.29	182	3.60	10	0.10	0.93
12/1/00 20:20	Test	1.47	9255	4.96	138	8.27	182	3.60	8	0.08	0.92
12/1/00 23:40	Train	1.47	9158	4.50	134	8.27	184	3.59	8	0.08	0.87
12/2/00 5:40	Test	1.47	8828	4.50	134	8.25	184	3.59	5	0.05	0.83
12/2/00 11:40	Train	1.47	7643	4.50	132	8.27	182	3.59	36	0.36	0.95
12/2/00 13:20	Train	1.47	7143	4.50	132	8.25	182	3.59	42	0.42	0.94
12/2/00 16:40	Train	1.47	8889	3.80	132	8.25	182	3.58	42	0.42	0.90
12/2/00 17:40	Test	1.47	10574	4.10	132	8.25	186	3.58	42	0.42	0.91
12/2/00 19:40	Train	1.47	13895	4.10	132	8.25	186	3.58	23	0.23	1.09
12/3/00 0:40	Train	1.47	15324	4.10	132	8.25	186	3.58	24	0.24	1.37
12/3/00 4:40	Train	1.47	13199	4.10	132	8.25	186	3.58	24	0.24	1.27
12/3/00 7:20	Prod.	1.47	11575	4.10	132	8.25	186	3.58	24	0.25	1.20
12/3/00 10:20	Train	1.47	10415	3.60	132	8.23	184	3.58	25	0.25	1.17
12/4/00 7:20	Train	1.46	11636	4.80	130	8.14	182	3.58	28	0.28	1.27
12/5/00 11:40	Test	1.51	10049	4.50	128	8.22	176	2.41	15	0.46	0.85
12/5/00 13:40	Train	1.59	11612	4.50	128	8.22	176	2.41	37	0.38	0.81
12/5/00 15:40	Train	1.61	12454	3.90	128	8.22	176	2.41	37	0.38	0.76
12/5/00 19:40	Train	1.62	12247	3.90	128	8.22	176	2.41	54	0.12	1.30
12/5/00 20:20	Prod.	1.62	12100	3.90	128	8.22	176	2.41	50	0.41	1.32
12/5/00 23:40	Train	1.62	11099	4.10	126	8.22	174	2.41	50	0.41	0.72
12/6/00 3:40	Train	1.62	10049	4.10	126	8.22	174	2.41	5	0.09	0.88
12/6/00 5:20	Train	1.62	9695	4.10	126	8.22	174	2.41	5	0.50	0.89
12/6/00 7:40	Train	1.62	9011	4.10	126	8.22	174	2.41	5	0.50	0.80
12/6/00 8:20	Test	1.62	8901	4.10	126	8.22	174	2.41	51	0.19	0.80
12/11/00 15:40	Train	1.66	8657	4.00	132	8.26	180	2.62	40	0.33	0.50
12/11/00 19:40	Test	1.72	7741	4.00	132	8.26	180	2.62	54	0.39	0.49
12/11/00 23:40	Prod.	1.74	6386	1.40	134	8.24	180	2.62	48	0.13	0.48
12/12/00 3:20	Prod.	1.74	5885	1.40	134	8.24	180	2.62	48	0.15	0.46
12/12/00 7:40	Train	1.75	5885	1.40	134	8.26	180	2.62	30	0.19	0.44
12/12/00 11:40	Train	1.73	6313	1.40	134	8.24	180	2.60	53	0.47	0.43
12/12/00 15:40	Train	1.77	9438	1.40	134	8.25	180	2.60	53	0.22	0.49
12/12/00 19:40	Train	1.76	7216	1.40	134	8.27	180	2.60	17	0.17	0.48
12/12/00 23:40	Train	1.77	6142	4.20	136	8.25	182	2.60	20	0.44	0.46
12/13/00 3:40	Prod.	1.78	5311	4.20	136	8.25	182	2.60	30	0.46	0.45
12/13/00 7:40	Prod.	1.80	4359	4.20	136	8.25	182	2.60	5	0.53	0.43

Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3 (cont.)

Date and Time	Set	Inf. Temp. (deg. C)	Inf. P.C. (counts/mL)	Inf. Colour (TCU)	Inf. Alk. (mg/L)	Inf. PH	Inf. Hard. (mg/L)	Flow (m3/h)	Alum (mg/L)	Polymer (mg/L)	Eff. Turb. (NTU)
12/13/00 11:40	Train	1.87	4664	5.30	133	8.25	186	2.60	14	0.44	0.44
12/13/00 15:20	Train	1.76	4286	5.30	133	8.23	186	2.61	25	0.28	0.37
12/13/00 19:40	Test	1.79	4725	5.30	133	8.23	186	2.61	47	0.19	0.39
12/13/00 23:40	Test	1.81	4530	4.68	138	8.25	190	2.61	47	0.37	0.39
12/14/00 3:40	Train	1.82	5531	4.68	138	8.25	190	2.61	28	0.16	0.41
12/14/00 7:40	Train	1.81	8217	4.68	138	8.23	190	2.61	15	0.46	0.46
12/14/00 11:40	Test	1.82	11416	5.30	148	8.24	208	2.61	22	0.20	0.56
12/14/00 15:40	Prod.	1.83	11221	5.10	148	8.25	208	2.61	13	0.54	0.68
12/14/00 19:40	Prod.	1.83	9670	5.10	148	8.25	208	2.61	17	0.22	0.67
12/15/00 9:40	Train	1.81	6471	5.80	164	8.23	208	2.61	34	0.42	0.73
12/15/00 12:00	Prod.	1.81	6288	5.80	164	8.23	208	2.61	55	0.44	0.73
12/15/00 13:20	Test	1.83	6288	5.80	164	8.23	208	2.61	55	0.44	0.73
12/15/00 19:20	Train	1.59	4762	5.80	164	8.24	208	2.66	55	0.55	0.77
12/16/00 10:20	Train	1.68	6068	5.80	161	8.25	214	2.66	55	0.55	0.73
12/16/00 15:00	Prod.	1.68	5763	5.80	161	8.23	214	2.66	55	0.55	0.70
12/16/00 18:00	Prod.	1.68	5189	5.80	161	8.23	214	2.66	55	0.55	0.68
12/17/00 3:40	Train	1.73	4713	5.80	161	8.23	214	2.66	55	0.55	0.53
12/17/00 11:00	Train	1.75	4737	5.50	154	8.21	214	2.66	55	0.55	0.60
12/17/00 16:40	Test	1.76	4261	5.50	154	8.19	214	2.66	55	0.55	0.68
12/18/00 1:00	Train	1.81	3553	5.50	154	8.19	214	2.66	55	0.55	0.72
12/18/00 10:00	Prod.	1.85	3101	5.50	154	8.19	214	2.66	5	0.28	0.67
12/18/00 17:40	Train	1.86	3150	5.70	158	8.01	208	2.66	5	0.31	0.61
12/19/00 6:00	Test	1.89	2955	6.10	160	8.01	206	2.66	34	0.09	0.54
12/19/00 11:40	Prod.	1.88	4615	6.80	156	8.01	204	2.66	5	0.26	0.48
12/20/00 13:20	Train	1.64	2576	6.90	154	8.00	200	3.47	56	0.05	0.44
12/20/00 21:20	Train	1.65	3211	6.60	164	8.00	202	3.44	55	0.56	0.40
12/21/00 3:20	Train	1.66	3162	6.60	164	8.00	202	3.43	55	0.55	0.41
12/21/00 8:40	Train	1.68	3748	6.60	164	8.02	202	3.48	55	0.55	0.39
1/8/01 20:00	Test	1.71	7534	5.70	142	8.01	184	3.00	5	0.55	0.65
1/9/01 8:00	Test	1.71	6679	6.00	140	8.01	180	3.00	5	0.55	0.54
1/10/01 8:00	Test	1.70	6532	5.90	138	7.99	180	3.07	5	0.55	0.42
1/10/01 13:40	Test	1.69	6593	5.40	138	7.99	180	3.00	5	0.55	0.41
1/10/01 17:40	Train	1.70	6667	4.90	138	7.98	178	3.01	26	0.07	0.41
1/10/01 21:40	Test	1.69	5714	4.90	138	7.98	180	3.02	39	0.42	0.39
1/11/01 5:40	Train	1.67	6532	4.90	138	7.98	180	2.97	20	0.13	0.38
1/11/01 9:40	Prod.	1.66	6642	4.50	136	7.98	178	3.01	27	0.07	0.39
1/11/01 13:40	Train	1.66	6484	4.80	136	7.98	178	3.01	10	0.48	0.39
1/11/01 17:40	Train	1.67	6252	4.80	136	7.98	178	3.01	29	0.14	0.38
1/11/01 21:40	Test	1.67	5421	5.60	136	7.98	178	3.03	45	0.34	0.36
1/12/01 1:40	Train	1.66	4945	6.10	136	7.98	180	3.02	19	0.40	0.35
1/12/01 7:20	Test	1.66	5482	6.10	136	7.98	180	3.00	36	0.26	0.33
1/12/01 9:00	Test	1.66	6007	4.40	130	7.98	176	3.00	36	0.26	0.34
1/12/01 13:40	Prod.	1.65	6801	5.10	130	7.99	176	3.02	48	0.54	0.37
1/12/01 17:20	Test	1.63	7338	5.10	130	7.99	178	3.01	40	0.55	0.42
1/12/01 21:20	Train	1.66	8230	4.40	130	7.99	180	3.01	50	0.55	1.13
1/13/01 7:20	Train	1.62	9988	4.40	130	7.99	180	3.00	46	0.55	0.71
1/13/01 14:40	Prod.	1.60	8364	4.40	130	7.99	180	3.02	47	0.55	0.68
1/15/01 6:00	Train	1.60	9805	4.30	128	8.03	186	2.99	48	0.55	0.71
1/15/01 11:20	Train	1.60	8938	7.23	142	8.02	180	3.00	12	0.30	0.89

Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3 (cont.)

Date and Time	Set	Inf. Temp. (deg. C)	Inf. P.C. (counts/mL)	Inf. Colour (TCU)	Inf. Alk. (mg/L)	Inf. pH	Inf. Hard. (mg/L)	Flow (m3/h)	Alum (mg/L)	Polymer (mg/L)	Eff. Turb. (NTU)
1/15/01 15:20	Train	1.57	8339	5.24	142	8.02	180	3.03	26	0.30	0.76
1/15/01 23:00	Train	1.58	6935	5.24	142	8.02	180	3.00	26	0.31	0.88
1/16/01 8:20	Train	1.58	8120	5.24	142	8.02	182	3.00	22	0.30	0.83
1/16/01 13:40	Train	1.59	7631	4.00	140	8.02	180	3.00	34	0.30	1.01
1/16/01 16:20	Prod.	1.59	7643	5.10	140	8.02	182	3.01	29	0.30	0.91
1/16/01 20:20	Test	1.60	7094	4.70	140	8.01	182	3.00	32	0.30	0.94
1/17/01 2:20	Test	1.61	6886	4.70	140	8.01	182	2.96	33	0.31	0.91
1/17/01 11:40	Test	1.61	8168	7.40	138	8.01	178	3.00	11	0.30	0.67
1/17/01 15:00	Train	1.62	8913	4.40	138	8.01	180	3.01	33	0.30	1.63
1/17/01 17:40	Prod.	1.62	9841	4.40	138	8.01	180	3.00	33	0.30	2.28
1/17/01 22:00	Train	1.60	11465	4.40	139	8.01	178	3.09	31	0.29	2.34
1/18/01 2:40	Train	1.59	14042	4.40	139	8.01	178	2.97	33	0.30	2.49
1/18/01 6:40	Train	1.58	14359	4.40	139	8.01	178	2.97	32	0.30	2.66
1/18/01 15:40	Prod.	1.73	12173	4.44	140	8.01	178	2.00	36	0.30	2.11
1/19/01 10:00	Train	1.73	14359	3.80	136	8.01	180	1.90	20	0.15	4.13
1/19/01 13:20	Train	1.70	12906	4.40	136	8.01	180	1.94	17	0.15	3.56
1/19/01 15:40	Prod.	1.67	12369	4.40	136	8.02	180	2.51	25	0.24	3.66
1/19/01 19:40	Train	1.65	12295	4.40	136	8.02	182	2.53	11	0.09	3.25
1/19/01 23:40	Prod.	1.67	13358	5.50	138	8.01	182	2.48	9	0.18	3.25
1/20/01 3:40	Train	1.67	13748	5.50	138	8.01	182	2.52	23	0.08	3.94
1/20/01 7:40	Train	1.66	13053	5.50	138	8.01	182	2.41	17	0.16	3.50
1/20/01 11:40	Test	1.65	11905	5.50	138	8.01	182	2.48	15	0.08	3.21
1/20/01 15:40	Train	1.66	11722	5.50	138	8.01	182	2.49	22	0.20	3.32
1/20/01 19:40	Train	1.67	13040	4.60	132	8.01	180	2.49	22	0.23	3.41
1/21/01 0:00	Train	1.67	15006	4.30	132	8.01	180	2.51	19	0.08	4.03
1/21/01 3:40	Train	1.66	15287	4.30	132	8.01	180	2.46	19	0.08	4.30
1/21/01 7:40	Train	1.66	14176	4.30	132	8.01	180	2.53	24	0.13	4.19
1/21/01 11:40	Prod.	1.64	12723	4.30	132	8.01	180	2.44	20	0.18	3.66
1/21/01 15:40	Test	1.64	12076	4.30	132	8.01	180	2.48	12	0.11	3.21
1/21/01 19:40	Train	1.66	12051	4.30	132	8.01	180	2.48	22	0.07	3.68
1/21/01 23:40	Train	1.66	11978	5.30	128	8.01	180	2.50	16	0.16	3.38
1/22/01 3:40	Test	1.66	11319	5.30	128	8.01	180	2.52	18	0.12	3.31
1/22/01 7:40	Train	1.66	10281	5.30	128	8.01	180	2.49	13	0.24	2.78
1/22/01 11:40	Prod.	1.79	9646	5.30	136	8.01	170	2.49	25	0.06	3.28
1/22/01 15:40	Prod.	1.85	9487	4.60	136	8.01	166	2.49	8	0.06	2.41
1/22/01 19:40	Train	1.84	9243	4.60	136	8.01	166	2.51	8	0.10	2.29
1/23/01 8:20	Train	1.87	9316	5.10	134	7.99	170	2.53	8	0.08	2.19
1/23/01 11:40	Train	1.79	8645	5.10	134	7.99	170	2.47	10	0.08	2.15
1/23/01 15:40	Prod.	1.85	8694	4.62	134	7.99	170	2.47	8	0.10	2.02
1/23/01 19:20	Train	1.85	8755	4.62	134	7.99	170	2.47	8	0.05	1.99
1/24/01 9:00	Test	1.83	10977	4.10	134	7.99	172	2.51	9	0.10	2.41
1/24/01 15:20	Train	1.84	10440	4.10	134	7.99	170	2.48	14	0.05	2.46
1/24/01 19:40	Train	1.85	11551	5.10	134	7.99	178	2.53	10	0.05	2.63
1/25/01 7:40	Train	1.85	12173	5.10	128	7.99	178	2.54	6	0.05	2.53
1/25/01 16:00	Test	1.93	13578	4.50	132	8.01	170	2.43	16	0.05	2.76
1/25/01 18:20	Train	1.86	15971	4.50	132	8.01	170	2.49	16	0.05	3.27
1/25/01 21:00	Prod.	1.86	17131	4.50	132	8.01	170	2.56	15	0.05	4.53
1/25/01 23:40	Train	1.86	17338	4.50	132	8.01	170	2.42	16	0.05	5.93
1/26/01 4:00	Test	1.85	17729	4.70	132	8.01	170	2.50	16	0.05	6.60

Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3 (cont.)

Date and Time	Set	Inf. Temp. (deg. C)	Inf. P.C. (counts/mL)	Inf. Colour (TCU)	Inf. Alk. (mg/L)	Inf. pH	Inf. Hard. (mg/L)	Flow (m3/h)	Alum (mg/L)	Polymer (mg/L)	Eff. Turb. (NTU)
1/26/01 8:40	Train	1.83	17411	4.70	128	8.01	172	2.49	16	0.05	5.82
1/26/01 14:40	Train	1.86	15726	4.70	128	8.01	174	2.50	16	0.05	4.91
1/26/01 23:00	Prod.	1.88	15348	7.41	130	8.01	174	2.51	16	0.05	4.43
1/27/01 4:40	Train	1.88	14799	7.41	130	8.01	174	2.51	16	0.05	4.03
1/27/01 10:40	Train	1.87	12772	7.41	130	8.01	174	2.52	16	0.05	3.44
1/27/01 17:20	Prod.	1.89	11783	4.60	134	8.03	176	2.49	16	0.05	3.07
1/27/01 23:40	Train	1.88	12308	5.24	136	8.03	174	2.47	16	0.05	3.13
1/28/01 9:20	Test	1.88	10769	4.70	136	8.01	180	2.51	16	0.05	2.83
1/28/01 17:40	Test	1.89	12063	4.70	136	8.01	176	2.48	16	0.05	2.71
1/29/01 4:00	Train	1.89	13114	4.30	138	8.01	176	2.54	16	0.05	3.03
1/29/01 9:00	Train	1.89	11673	4.80	124	8.01	178	2.50	16	0.05	2.82
1/29/01 11:00	Train	1.84	11087	4.80	124	8.01	178	3.23	8	0.16	3.04
1/31/01 13:00	Train	1.80	10391	3.70	132	7.98	168	2.98	17	0.04	2.35
1/31/01 16:20	Prod.	1.82	11575	3.30	132	7.98	168	3.01	20	0.08	2.69
1/31/01 20:00	Train	1.82	13480	3.30	128	7.98	164	2.94	20	0.08	2.98
1/31/01 23:40	Prod.	1.82	14005	3.30	128	7.98	164	3.01	20	0.08	3.31
2/1/01 3:20	Train	1.81	13614	3.30	128	7.98	164	3.01	20	0.08	3.30
2/1/01 6:20	Train	1.81	12833	3.30	128	7.98	164	3.00	20	0.08	3.12
2/1/01 8:20	Train	1.81	12283	3.60	128	7.98	166	2.98	15	0.08	2.99
2/2/01 11:20	Prod.	1.74	12100	3.02	122	8.01	166	2.98	16	0.08	3.02
2/6/01 14:00	Train	1.47	9316	4.34	124	8.03	168	2.76	54	0.32	1.79
2/7/01 12:20	Prod.	1.45	9316	3.60	126	7.99	166	2.52	55	0.33	1.68
2/9/01 13:40	Train	1.50	8767	3.49	124	8.03	170	2.24	55	0.32	1.49
2/12/01 21:40	Prod.	1.56	11770	6.90	130	8.06	174	2.27	54	0.32	1.40
2/13/01 12:20	Train	1.54	9390	4.30	130	8.05	174	2.25	55	0.33	1.36
2/16/01 8:40	Prod.	1.60	11074	2.70	136	8.02	174	1.98	38	0.25	1.51
2/16/01 13:00	Test	1.60	10220	2.70	136	8.02	174	1.99	42	0.29	1.38
2/16/01 17:00	Train	1.62	9719	3.80	136	8.02	182	2.03	21	0.12	2.01
2/17/01 13:00	Prod.	1.59	8510	3.70	128	8.02	178	2.04	37	0.16	2.32
2/17/01 17:00	Prod.	1.61	8767	3.70	128	8.02	178	2.01	18	0.31	1.25
2/17/01 21:00	Prod.	1.61	10317	3.70	128	8.02	178	1.99	40	0.23	1.62
2/18/01 1:00	Train	1.61	9817	4.10	130	8.02	176	2.04	24	0.14	2.03
2/18/01 9:00	Test	1.59	8840	4.10	130	8.02	176	1.96	30	0.07	2.62
2/18/01 13:00	Test	1.59	8767	4.10	130	8.02	176	1.99	48	0.27	1.59
2/18/01 17:00	Train	1.60	8791	3.54	128	8.02	178	2.00	23	0.30	1.16
2/18/01 21:00	Train	1.60	10513	3.40	126	8.02	178	1.92	31	0.11	2.35
2/19/01 1:00	Test	1.61	10440	3.40	126	8.02	178	1.98	40	0.21	1.95
2/19/01 5:00	Test	1.61	10501	3.40	126	8.02	178	2.01	29	0.27	1.28
2/19/01 9:00	Train	1.60	9573	3.40	126	8.00	178	1.99	20	0.27	1.22
2/19/01 13:00	Test	1.61	8962	3.20	128	8.00	178	1.95	35	0.16	2.13
2/19/01 21:00	Train	1.61	10488	3.50	132	8.00	174	1.96	34	0.24	1.52
2/20/01 1:00	Test	1.59	10745	3.50	132	8.00	174	1.98	13	0.07	2.14
2/20/01 5:00	Train	1.59	10659	2.90	132	8.00	172	2.06	13	0.16	1.86
2/20/01 13:00	Prod.	1.57	9341	3.40	124	8.00	174	1.97	11	0.16	1.66
2/20/01 17:00	Train	1.59	8889	3.20	124	8.00	174	1.98	6	0.24	1.43
2/20/01 21:00	Prod.	1.60	9792	3.54	124	8.00	176	1.99	13	0.13	1.77
2/21/01 5:00	Test	1.59	10317	3.54	124	8.00	176	2.00	10	0.30	1.45
2/21/01 9:00	Train	1.60	9890	3.54	123	8.00	176	1.99	45	0.22	1.94
2/23/01 18:00	Test	1.63	14347	4.80	134	8.04	178	3.00	35	0.32	2.23

Table A.1 Modelling data for EPCOR Pilot Plant control system, Model 3 (cont.)

Date and Time	Set	Inf. Temp. (deg. C)	Inf. P.C. (counts/mL)	Inf. Colour (TCU)	Inf. Alk. (mg/L)	Inf. pH	Inf. Hard. (mg/L)	Flow (m3/h)	Alum (mg/L)	Polymer (mg/L)	Eff. Turb. (NTU)
2/24/01 5:00	Prod.	1.54	12283	4.60	130	8.02	178	2.97	5	0.32	2.32
2/24/01 9:00	Train	1.53	11600	4.60	124	8.02	176	3.00	5	0.32	2.24
2/24/01 13:40	Test	1.52	11551	4.20	124	8.02	176	2.96	5	0.32	2.09
2/24/01 23:00	Prod.	1.47	12882	4.20	124	8.02	176	2.99	46	0.32	1.96
2/25/01 3:20	Train	1.48	12002	4.20	124	8.02	176	2.97	28	0.32	1.70
2/25/01 11:20	Train	1.46	10916	4.20	124	8.02	176	2.98	5	0.29	1.84
2/25/01 17:00	Train	1.47	11013	4.20	124	8.04	176	3.02	5	0.26	1.84
2/25/01 21:40	Train	1.48	11917	4.50	124	8.04	176	3.00	5	0.32	2.01
2/26/01 8:00	Train	1.46	9438	3.70	124	8.04	176	3.01	5	0.04	1.76

APPENDIX B – CURRICULUM VITAE

Christopher W. Baxter

Suite 1425, 5328 Calgary Trail South
Edmonton, Alberta
T6H 4J8
(780)903-9002
cwbaxter@hydrannt.com

Biographical Data

Date of Birth: July 22,1974
Place of Birth: Pointe-Claire, Quebec
Citizenship: Canadian Citizen

Education

Ph.D. Environmental Science (program complete, to be conferred June 6, 2002)
University of Alberta, Edmonton, Alberta
Department of Civil and Environmental Engineering

- 9.0 G.P.A. (9 point scale)

M.Sc. Environmental Science (11/98)
University of Alberta, Edmonton, Alberta
Department of Civil and Environmental Engineering

- 8.8 G.P.A. (9 point scale)

B.Sc.(Hon.) Environmental Science (Chemistry) (05/96)
Queen's University, Kingston, Ontario
Department of Chemistry

- Graduated with first-class honours

Honours and Awards

- Recipient of the 2001 Dr. Donald R. Stanley Scholarship in Environmental Engineering
- Recipient of a 2001 Martha Piper Award in Research Communication Excellence
- Recipient of a 2001 Mary Louise Imrie Graduate Student Award
- Recipient of a 2000 Izaak Walton Killam Memorial Scholarship
- Recipient of the 2000 Dr. Donald R. Stanley Scholarship in Environmental Engineering

- Recipient of a 1998 Natural Sciences and Engineering Research Council of Canada Industrial Postgraduate Scholarship
- Recipient of a 1998 Walter H. Johns Graduate Fellowship
- Recipient of the 1998 Bruce F. Willson Scholarship in Urban Engineering
- Recipient of a 1998 Martha Piper Award in Research Communication Excellence
- Recipient of a 1998 Graduate Studies and Research Scholarship
- Recipient of the 1997 Bruce F. Willson Scholarship in Urban Engineering
- Recipient of the 1995 Senator Frank Carrel Upper-Year Scholarship

Work Experience

HYDRANNT Consulting Services
Principal (05/01 – present)

- Develop and implement artificial intelligence models and process control application for clients in the drinking water treatment industry
- Assist utilities at developing in-house expertise in the use of artificial intelligence techniques
- Further the progress of artificial intelligence applications in the water treatment industry through continued research and development

Department of Civil and Environmental Engineering, University of Alberta
Graduate Student (09/96-present)
Supervisors: Stephen J. Stanley and Daniel W. Smith

- Design and implement the drinking water treatment industry's first artificial neural network model-based advanced process control system for the EPCOR Water Services pilot facility
- Design artificial intelligence models for water treatment facilities operated by the Metropolitan Water District of Southern California and implement the models in process control
- Design an artificial neural network model for the removal of natural organic matter by enhanced coagulation, and implement the model in process control at the Rossdale and E.L. Smith Water Treatment Plants in Edmonton
- Maintain effective lines of communication with project sponsors (American Water Works Association Research Foundation, and the Metropolitan Water District of Southern California) through periodic oral presentations and written reports
- Gain experience in the areas of water treatment, wastewater treatment, water quality management, solid waste management, and environmental chemistry through graduate-level course work

Zenon Environmental Laboratories, Anjou, Quebec
Laboratory Technician (05/94-08/94 and 05/95-08/95)

- Oversaw the preparation of soil, water, and air samples for the analysis of organic contaminants
- Developed a quality assurance monitoring program for all laboratory analyses

Publications and Presentations

Publications in Refereed Journals

Baxter, C.W., Q. Zhang, S.J. Stanley, R. Shariff, R-R. T. Tupas, and H.L. Stark. 2001. Drinking Water Quality and Treatment: The Use of Artificial Neural Networks. *Can. J. Civ. Eng.* 28(Suppl. 1): 26-35

Baxter, C.W., S.J. Stanley, and Q. Zhang. 1999.Development of a Full-Scale Artificial Neural Network Model for the Removal of Natural Organic Matter by Enhanced Coagulation. *J. Water SRT - AQUA*, 48(4):129-136.

Baxter, C.W., R. Shariff, S.J. Stanley, D.W. Smith, Q. Zhang, and E.D. Saumer. Model-based Advanced Process Control of Coagulation. *Wat. Sci. Tech.* (accepted 08/01).

Baxter, C.W., S.J. Stanley, Q. Zhang., D.W. Smith. Developing Artificial Neural Network Models of Drinking Water Treatment Processes: A Guide for Utilities. (submitted 08/01).

Other Refereed Publications

Baxter, C.W., Tupas, R-R.T., Zhang, Q., Shariff, R., Stanley, S.J., Coffey, B.M., and Graff, K.G. 2001. *Artificial Intelligence Systems for Water Treatment Plant Optimization.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 141 p.

Stanley, S.J., C.W. Baxter, Q. Zhang, and R. Shariff. 2000. *Process Modeling and Control of Enhanced Coagulation.* American Water Works Association Research Foundation and American Water Works Association, Denver, CO. 167 p

Chapters in Edited Books

Baxter, C.W., R. Shariff, S.J. Stanley, and Q. Zhang. Advanced Process Control with Online Monitors. In *Online Monitoring.* E. Hargesheimer, Ed. AWWA and ProAqua. (in press)

Publications in Conference Proceedings

Baxter, C.W., R. Shariff, S.J. Stanley, D.W. Smith, Q.Zhang, and E.D. Saumer. 2001. Model-based Advanced Process Control of Coagulation. In *Proceedings of the 1st IWA Conference on Instrumentation, Control, and Automation*. Malmo, Sweden: IWA.

Baxter, C.W., D.W. Smith, and S.J. Stanley. 2001. Using Artificial Neural Networks to Analyze Pilot-Scale Data. In *Proceedings of the 1st IWA Conference on Instrumentation, Control, and Automation*. Malmo, Sweden: IWA.

Baxter, C.W., S.J. Stanley, B.M. Coffey, and K.G. Graff. 2001. Feasibility of Artificial Intelligence Systems for Water Treatment Process Modeling and Control at Metropolitan Water District of Southern California Facilities. In *Proceedings of the 2001 Information Management and Technology Conference*. Atlanta, GA.: AWWA.

Baxter, C.W., S.J. Stanley, Q. Zhang, and D.W. Smith. 2000. Developing Artificial Neural Network Process Models: A Guide for Drinking Water Utilities. In *Proceedings of the 2000 Annual Conference of the Canadian Society of Civil Engineering*. London, ON.: CSCE.

Baxter, C.W., S.J. Stanley, and Q. Zhang. 1999. Applications of Full-Scale Artificial Neural Network Models of Enhanced Coagulation. In *Proceedings of the 1999 Information Management and Technology Conference*. New Orleans, LA.: AWWA.

Zhang. Q., C.W. Baxter, S.J. Stanley, and R. Shariff. 1999. Automating Water Treatment Plants Using Artificial Neural Network Models. In *Proceedings of the 1999 Engineering & Construction Conference*. Orlando, FL.: AWWA.

Stanley, S.J., Q. Zhang, and C.W. Baxter. 1998. Use of Artificial Neural Network Models in Water Treatment Applications. In *Proceedings of the Third Annual Research and Development Conference on Rural Sanitation*. Anchorage, AK.: Alaska Water and Wastewater Management Association.

Baxter, C.W., S.J. Stanley, and Q. Zhang. 1998. Development of a Full-Scale Artificial Neural Network Model for Enhanced Coagulation. In *Proceedings of the 50th Annual Conference of the Western Canada Water and Wastewater Association*. Calgary, AB.: WCWWA.

Theses

Baxter, C.W. 2002. Applications of Artificial Neural Networks in Drinking Water Treatment Process Modelling and Control. Doctoral thesis. University of Alberta, Edmonton, AB.

Baxter, C.W. 1998. Full-Scale Artificial Neural Network Modelling of Enhanced Coagulation. Master's thesis. University of Alberta, Edmonton, AB.

Consulting Reports

Baxter, C.W., D.W. Smith, T. Yu, and D. Bromley. 2001. Preliminary Investigation of a Radial-Flow Fluidizable Filter Unit for Drinking Water Filtration. A report prepared for Woodcock and Associates, Inc., Vancouver, BC. 41 p.

Baxter, C.W., and N.J. Harper. 1998. Evaluation of Proposed Benchmarking Partners for AQUALTA Inc. A report prepared for AQUALTA Inc., Edmonton, AB. 37 p.

Selected Presentations

October 28, 2001. Improving Water Utility Management Through the Use of Artificial Intelligence. Oral workshop presentation at the *2001 WCWWA Annual Conference.* Edmonton, AB: WCWWA.

June 17, 2001. Improving Water Utility Management Through the Use of Artificial Intelligence. Oral workshop presentation at the *2001 AWWA Annual Conference and Exhibition.* Washington, D.C.:AWWA.

June 4, 2001. Model-based Advanced Process Control of Coagulation. Oral presentation at the *1st IWA Conference on Instrumentation, Control, and Automation.* Malmo, Sweden: IWA.

June 4 – 6, 2001. Using Artificial Neural Networks to Analyze Pilot-Scale Data. Poster presentation at the *1st IWA Conference on Instrumentation, Control, and Automation.* Malmo, Sweden: IWA.

April 10, 2001. Feasibility of Artificial Intelligence Systems for Water Treatment Process Modeling and Control at Metropolitan Water District of Southern California Facilities. Oral presentation at the *2001 Information Management and Technology Conference.* Seattle, WA: AWWA.

March 15, 2000. Intelligent Automation of Water Treatment Processes. Poster presentation at the Graduate Students' Poster Night. Canadian Society of Civil Engineering. Edmonton, AB

March 9, 2000. Intelligent Automation of Water Treatment Processes. Poster presentation at the Graduate Students' Association Awards Night. University of Alberta. Edmonton, AB.

June 7, 2000. ANN Model Building and Applications. Oral workshop presentation to EPCOR Water Services. Edmonton, AB.

May 9, 2000. ANN Model Building and Applications. Oral workshop presentation to the Metropolitan Water District of Southern California. Laverne, CA.

April 16, 2000. The ANN Model Building Process. Oral workshop presentation at the *2000 Information Management and Technology Conference*. Seattle, WA: AWWA.

April 16, 2000. ANN Model Applications. Oral workshop presentation at the *2000 Information Management and Technology Conference*. Seattle, WA: AWWA.

April 5, 2000. Drinking Water in Edmonton and Around the World. Oral presentation at *Science Week 2000*. J.H. Picard H.S. Edmonton, AB.

March 16, 2000. Water Treatment Process Control Using Artificial Neural Network Models. Poster presentation at the Graduate Students' Poster Night. Canadian Society of Civil Engineering. Edmonton, AB

March 15, 2000. Water Treatment Process Control Using Artificial Neural Network Models. Poster presentation at the Graduate Students' Association Awards Night. University of Alberta. Edmonton, AB.

February 24, 2000. Waste and Our World. Interactive presentation for local elementary school students. University of Alberta Outreach Program. Edmonton, AB.

February 5, 2000. Applications of Artificial Neural Network Models in Water Treatment Process Control. Poster presentation at *Research Revelations 2000*. University of Alberta. Edmonton, AB.

April 19, 1999. Applications of Full-Scale Artificial Neural Network Models of Enhanced Coagulation. Oral presentation at the *1999 Information Management and Technology Conference*. New Orleans, LA.: AWWA.

October 27, 1998. Development of a Full-Scale Artificial Neural Network Model for Enhanced Coagulation. Oral presentation at the *50th Annual Conference of the Western Canada Water and Wastewater Association*. Calgary, AB.: WCWWA.

February 9, 1998. Development of a Full-Scale ANN Model for the Removal of NOM by Enhanced Coagulation. Presented at the *Central Canadian Water and Wastewater Conference*. Burlington, ON: CAWQ.

Teaching Experience

Department of Civil and Environmental Engineering, University of Alberta
Principal Instructor, CIV E 729 (Fall 2001)

- Instruct graduate students in the application of the artificial neural network technology to the field of environmental engineering
- Prepared all course materials, assignments, and tutorial sessions
- Developed and maintained a course website with useful course material

Department of Civil and Environmental Engineering, University of Alberta
Principal Instructor, CIV E 290 (Winter 2000 and Winter 2001)

- Developed a teaching plan, lecture materials, assignments, and exams for 120 students enrolled in the department's compulsory 2nd year statistics course
- Developed and maintained a course website where students could obtain assignment and exam answers, as well as additional statistics resources
- Received a mean score of 4.0/5.0 on the instructor-specific questions of the university's student course evaluations

Department of Civil and Environmental Engineering, University of Alberta
Guest Lecturer, CIV E 220 (Winter 1999)
Course Instructor: Stephen J. Stanley

- Presented a lecture on the theory and applications of coagulation and softening in drinking water treatment

Department of Civil and Environmental Engineering, University of Alberta
Course Marker, ENV E 220 (Fall 1998)
Course Instructor: Warren B. Kindzierski

- Assisted students in the development of sound scientific writing skills through the grading of term papers and abstracts and reviewing my comments with the students

Department of Civil and Environmental Engineering, University of Alberta
Guest Lecturer, CIV E 321 (Winter 1998)
Course Instructor: Stephen J. Stanley

- Presented lectures on risk assessment and contaminant modelling

Department of Civil and Environmental Engineering, University of Alberta
Laboratory Instructor, ENV E 220 (Fall 1997)
Course Instructors: Warren B. Kindzierski and Ian D. Buchanan

- Provided instruction on the theoretical and practical aspects of the laboratory component of the course
- Assisted students in the development of sound laboratory and writing skills through the grading of laboratory reports and maintaining office hours for individual consultation

Association Memberships

- American Water Works Association
- Western Canada Water and Wastewater Association

Professional Service and Volunteer Work

- Assist in the application of engineering technologies to developing world problems as Director, Water Research with Engineers Without Borders Canada and as VP Resources of the University of Alberta Chapter of Engineers Without Borders
- Appointed secretary of the Information Technology and Management Research Committee, Water Science and Research Division of the American Water Works Association
- Educate local elementary and high school students on drinking water treatment and conservation through interactive presentations with the University of Alberta's Outreach Program
- Maintain effective lines of communication between faculty members and graduate students as the founder and chair of the Environmental Engineering Graduate Students' Society at the University of Alberta
- Represented the University of Alberta as a Judge, Team Leader, and Tour Volunteer at the 1999 Canada Wide Science Fair in Edmonton, Alberta
- Promote the social, physical, and mental development of youth by organizing weekly meetings, camping trips, and other outdoor activities as the leader of a local Scout group.

Graduate Course Work

- CIV E 620: Environmental Engineering Measurements I
- CIV E 622: Physical/Chemical Water and Wastewater Treatment
- CIV E 624: Biological Waste Treatment Processes
- CIV E 625: Engineering Management of Water Quality
- CIV E 626: Environmental Health Engineering
- CIV E 627: Environmental Engineering Measurements II
- CIV E 628: Municipal Solid Waste Management
- CIV E 658: Design of Civil Engineering Experiments
- CIV E 729: Artificial Neural Networks

Professional References

S.J. Stanley, Ph.D., P.Eng.
General Manager, Water Services
EPCOR Water Services
10065 Jasper Avenue
Edmonton, AB, T5J 3B1
(780) 412-7755

D.W. Smith, Ph.D., P.Eng.
Professor, Department of Civil and Environmental Engineering
Room 304 EEB, University of Alberta
Edmonton, AB, T6G 2M8
(780) 492-4138

I.D. Buchanan, Ph.D., P.Eng.
Associate Professor, Department of Civil and Environmental Engineering
Room 304 EEB, University of Alberta
Edmonton, AB, T6G 2M8
(780) 492-0244